

Everybody's fault

The Kashmiri earthquake highlights the urgent need for Pakistan and India to put aside their differences and build stronger scientific ties.

In the past 15 years, the dispute between India and Pakistan over the military 'line of control' that divides Kashmir may have cost between 40,000 and 60,000 lives. But even that formidable death toll was surpassed in a few moments last autumn, when the boundary between the Indian and Eurasian tectonic plates shook, killing an estimated 75,000 people. The earthquake of magnitude 7.6 left many more homeless on both sides of the border.

The mayhem was made worse by the failure of either Pakistan or India to adequately address their well-established earthquake vulnerabilities. As a News Feature on page 16 of this issue reports, there is now a movement among Pakistan's geologists to deal with earthquakes. But so far, there has been no real move towards closer collaboration with researchers in neighbouring India.

Such collaboration has the potential to significantly strengthen the two nations' earthquake defences, potentially saving thousands of lives. Kashmir is home to some of the subcontinent's poorest people. They cannot afford reinforced concrete or 'seismic isolation', but they can make simple modifications to house designs — such as integrating a roof into a house's frame — that will greatly improve their chances of surviving future earthquakes. Government facilities such as schools, bridges and hospitals could also be built at relatively low cost with earthquakes in mind, dramatically strengthening future disaster response in the region.

Real change will still require substantial public investment, however. Better scientific cooperation could help both nations set priorities for such investment. By mapping and studying the many faults that run through the line of control, Pakistani and Indian seismologists could build a better understanding of which towns are at greatest earthquake risk. Together they could devise building regulations that would be sensitive to the region's limited resources while ensuring that more of their citizens would survive. They could also provide insight into how relief could be most efficiently distributed to the region in the aftermath of a major earthquake.

Both countries have recently tried to de-escalate the conflict over

the disputed province of Kashmir. Collaboration in all fields of civil society, including science, has a role to play here too (see *Nature* 393, 499; 1998). There are finally some promising signs: in the past few years, scientific leaders from the nations have met, and joint efforts are now under way in fields such as agricultural biotechnology.

But when it comes to seismology in Kashmir, the two nations are still barely talking. They are both paranoid about ceding military advantage in the disputed region — with the result that geologists bearing maps or GPS receivers are banned from entry. Consorting with 'enemy' researchers is also viewed with deep suspicion.

In May, however, geologists from both sides of the border will attend a 'science for peace' workshop in Lahore, Pakistan, where they will discuss topics such as remote sensing, geophysics and seismology, as well as the idea of turning the hotly contested Nanga Parbat glacier into a scientific peace park. The meeting's US coordinators hope that it can serve as a starting point for scientific exchange in geology.

Both nations' scientists should take advantage of this gathering to begin serious and sustained collaboration. Their governments should make provisions to allow scientific exchange and the open dissemination of data of the sort that is needed for such work. Indian and Pakistani scientists based abroad, who often work side-by-side in Western laboratories, can help by acting as go-betweens for researchers in the region who currently share no direct contact.

If India and Pakistan are to get to grips with earthquakes, they must overcome enormous technical, social and economic challenges. The task is formidable and requires the two nations to put their manifest political differences to one side and work together. By allowing their scientists to collaborate openly, governments in both countries can gain an understanding of the risks their citizens face, and help them prepare for the future.

"When it comes to seismology in Kashmir, the two nations are still barely talking."

Evaluate this

The objective evaluation of research isn't working as it should.

The rigorous evaluation of research projects and programmes is in increasingly common demand across the world. Attempts have been made to implement it in Europe, Japan and the United States — but until the calibre of these efforts improves, scientists will continue, justifiably, to view them with suspicion.

Policy-makers have talked for years about the need to rigorously

evaluate research programmes that consume billions of dollars of taxpayers' money. Researchers — especially those doing basic research that can't be readily tied to concrete outcomes — have tended to be sceptical. Nonetheless, evaluation is now under way on a significant scale in every major economy.

Yet nowhere is the circle between research programmes, evaluation and research funding decisions quite complete. A process for measuring 'performance' is firmly in place at many agencies. Yet few research managers genuinely believe that the outcomes of these assessments are really driving funding decisions.

Take the United States, where the Government Performance and Results Act demands significant qualitative assessment of all federal



programmes. The Bush administration has also proactively pursued research evaluation through the all-powerful White House Office of Management and Budget (OMB), which sets the president's annual budget proposal. A lot of evaluation is now taking place. And to its credit, the OMB seems to have focused much of its attention on programmes that are not properly peer reviewed.

But in the end, what is the evidence that anyone in the government is listening? Where are the examples of programmes that the administration doesn't like being revived because they perform well — or ones that it intuitively favours being cut back?

Evaluation that doesn't work is, according to a discussion at the annual meeting of the American Association for the Advancement of Science (AAAS) last month, worse than none at all. It costs a substantial amount of money — anything from 0.25% to 2% of the cost of the programme under scrutiny — and it exhausts and sometimes

demoralizes the researchers obliged to participate. Some argue that this process is of inherent value in lending direction to projects and programmes, but that is a minority view.

Even so, demands for accountability will not go away. The systems in place, flawed as they may be, are unlikely to be dismantled. In Japan, hard-done-by researchers are in rebellious mood (see *Nature* **438**, 1051–1052; 2005). And in the United States, the process contains little of the transparency Americans expect of their government.

The OMB is almost as secretive as it is powerful — but researchers need to be convinced that all of its evaluation is leading somewhere. Congress should therefore ask the Government Accountability Office to report on the OMB evaluation process as it relates to science and technology. Such a study, from a watchdog of established integrity, might reassure research managers that they are not being sent each year on a wild goose chase. ■

Gradual force

A delicate probe, twenty years old this week, has transformed our understanding of the nanoscale.

What difference can a breakthrough in science make over two decades? A quick comparison of the respective fates of two discoveries made twenty years ago reaffirms how daft it is to try to predict research outcomes over such a timescale.

Both tales begin at IBM's Zurich research laboratory in the early 1980s. In one corner of the lab, Gerd Binnig, Heinrich Rohrer and others were building an instrument that would come to be known as a scanning tunnelling microscope (STM). In another, Georg Bednorz and Alexander Müller were doing experiments on materials that they thought might hold promise as superconductors.

In early 1986, Bednorz and Müller got the first hints that they were on to something. An oxide of lanthanum, barium and copper seemed to retain the ability to conduct electricity without resistance at temperatures of up to 35 K (−238 °C). This, at the time, was striking: for more than a decade, the ceiling on transition temperatures for superconductors had been stuck at less than 24 K.

Within months of publishing their first paper (*Z. Phys. B* **64**, 189–193; 1986), the result was successfully replicated, and interest in its ramifications exploded. The buzz over high-temperature superconductivity at the March 1987 meeting of the American Physical Society in New York was such that *The New York Times* dubbed the event “the Woodstock of physics”.

By the time Bednorz and Müller picked up their Nobel prize in December of the same year, similar ceramic materials that could superconduct at the temperature of liquid nitrogen, 77 K (−196 °C), had been discovered. This, it was thought, would open the door to widespread practical use — and, briefly, industrial and government funding surged in at the speed of the magnetically levitating trains that the field was, according to countless news stories, due to produce in the fullness of time.

By contrast, few members of the public had even heard of the humble scanning tunnelling microscope, which Binnig and Rohrer first described in an internal IBM document in March 1981. It took

until 1986, by which time the STM was already widely used to study materials, for them to receive the Nobel prize.

An STM exploits a quantum phenomenon called electron tunnelling: when two conducting materials are held close together and a voltage is applied, electrons hop from one to the other, producing a current that is highly sensitive to the distance of separation. An STM scans a sharp tip over a surface, translating the current registered into a topographical map of the surface. The instrument was powerful enough to reveal the positions of atoms.

But the STM only worked for materials that conduct electricity, which many of the things that interest biologists and materials scientists don't do. This problem was addressed by a paper published twenty years ago this week, on 3 March 1986. The paper, ‘Atomic force microscope’ (*Phys. Rev. Lett.* **56**, 930–933; 1986), introduced the instrument that has fuelled the current explosion of interest in nanotechnology.

The atomic force microscope traces topography by scanning a sharp tip over the surface. It measures tiny forces between the tip and the sample via the deflection of a thin cantilever to which the tip is attached. This is an immensely versatile technique (see page 14), and that first paper has clocked up more than 4,500 citations.

High-temperature superconductivity, on the other hand, never quite lived up to its commercial hype. The materials in question have been difficult to fabricate, their properties are more constrained than some physicists had hoped, and there is still no agreement on why they work as they do (see papers in this month's *Nature Physics*). The materials remain rich systems for experimental study and pose intriguing theoretical problems, but their practical use is largely confined to making superconducting quantum interference devices (SQUIDS) to measure magnetic fields, and prototype transmission lines that can carry high-density current.

In terms of their technological and economic application, it looks like the dark horse made it to the wire first. That should serve to remind managers of research agencies and industrial laboratories of the folly of trying to predict how much value any particular scientific breakthrough may hold. ■

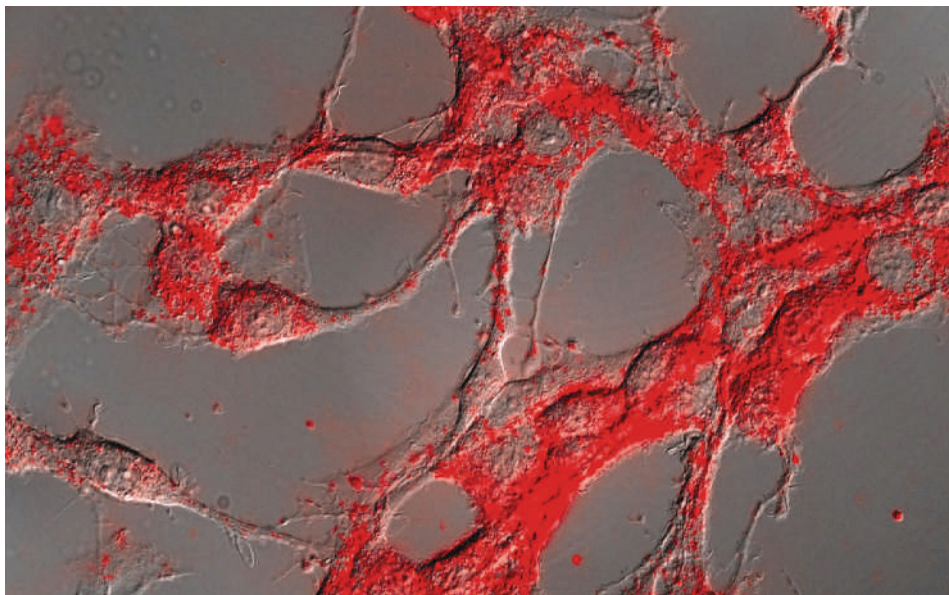
“Atomic force microscopy is an immensely versatile technique.”

RESEARCH HIGHLIGHTS

A bright idea

Nature Biotechnol. doi:10.1038/nbt1188 (2006)
Researchers in the United States have designed a new kind of glowing nanocrystal, or quantum dot, for *in vivo* biological imaging.

Quantum dots can be targeted at certain tissues or processes. But their usefulness in imaging has been limited by the need for external illumination to make the dots glow. Light can't penetrate deep into tissue, and the dots' visibility is reduced by the tissue itself fluorescing. Now Jianghong Rao and his colleagues at Stanford University, California, suggest a solution. They linked dots to the light-producing enzyme luciferase. When the luciferase is turned on by the addition of a substrate, its light makes the quantum dots glow — pictured here in glioma cells, a type of neural cell.



CANCER BIOLOGY

Can't resist

Cancer Cell 9, 133–146 (2006)

Researchers have found a way to tackle the resistance that some cancers show to certain kinds of chemotherapy.

Brent Stockwell and his team at Columbia University, New York, focused on the commonly used cancer drug doxorubicin, which triggers cancer cells to commit suicide through a process known as apoptosis. However, in many cancers the cell's apoptosis machinery is crippled.

By screening a range of compounds, the team identified one group of compounds, known as indoxins, which can make such cells more responsive to the drug. Indoxins increased the expression of the protein needed for doxorubicin to trigger cell death, and stalled the cells at an especially sensitive point in their division cycle.

ASTRONOMY

Drifting clouds

Astrophys. J. 638, 772–785 (2006)

The Large Magellanic Cloud, a dwarf galaxy that orbits the Milky Way, has been clocked moving at 378 ± 18 kilometres per second relative to our own Galaxy.

The measurement, made using the Hubble Space Telescope to track the galaxy's movement in relation to 21 distant quasars, is by far the most accurate until now.

It should help astronomers to deduce the size, shape and density of the dark-matter halo that constitutes most of the mass in our

Galaxy and which keeps the Large Magellanic Cloud in orbit, say Nitya Kallivayalil of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, and her colleagues.

CHEMISTRY

Supersize me

Angew. Chem. Int. Edn doi:10.1002/anie.200504047 (2006)

The discovery in the early 1950s that metals could be 'sandwiched' between organic molecules led to tremendous advances in organometallic chemistry, producing useful chemicals such as magnetic resonance imaging (MRI) contrast agents and industrial catalysts. Now chemists have supersized what was the very first such sandwich, a compound known as ferrocene (Cp_2Fe).

In ferrocene, two ring-shaped molecules — cyclopentadienyls (Cp) — serve as bread

around an iron-atom filling. Peter Vollhardt of the University of California, Berkeley, and his colleagues supersized the bread by attaching a ferrocene to each of the five carbon atoms in another Cp ring. Sandwich compounds made from this molecule (pictured below) may have unusual magnetic or electrochemical properties.

SOLID-STATE PHYSICS

Mind the step

Nature Phys. doi:10.1038/nphys245 (2006)

Quantum physics does strange things to electrons in graphite. Graphene — a single sheet of graphitic carbon — differs from other semiconducting materials in how it manifests the quantum Hall effect, whereby step-like changes are seen in a material's conductivity as a magnetic field is applied.

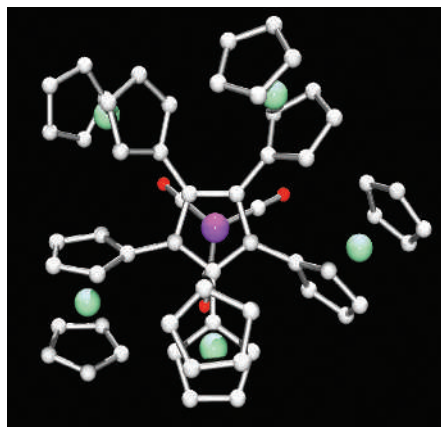
Andre Geim of the University of Manchester, UK, and his co-workers showed recently that the conductivity steps sit at different values in graphene compared with other thin films (*Nature* 438, 197–200; 2005). Now they have found another new form of the quantum Hall effect, in double layers of graphene. It is marked by the disappearance of the lowest 'step', and by metal-like rather than insulating behaviour in high magnetic fields.

CELL BIOLOGY

Uneven split

Science 311, 1146–1149 (2006)

Scientists have long hypothesized that stem cells retain their oldest DNA when they divide, to avoid accumulating replication



errors — but this has been difficult to prove.

Now Athanasios Armakolas and Amar Klar, both at the National Cancer Institute at Frederick, Maryland, have used a neat genetic trick to track the inheritance of DNA in dividing cells. The two strands in a DNA double helix are never the same age. One, call it Crick, will have been the template against which the other, Watson, was built. The researchers find that in certain embryonic mouse cells, including embryonic stem cells, the replication of the DNA in chromosomes is handled so that copies containing the old Cricks end up in one daughter cell whereas copies containing the old Watsons end up in the other. In other types of cell the inheritance was random.

IMMUNOLOGY

Complementary effects

EMBO J. doi:10.1038/sj.emboj.7601004 (2006)
The proteins of the mammalian complement system help to protect against infectious disease by promoting inflammation. Local activation of the complement system in the brain after stroke, however, is thought to contribute to brain damage.

Marcela Pekna from Göteborg University in Sweden and her colleagues have now discovered a more positive side to the story. They show that complement proteins also promote the maturation of replacement neural cells in both normal situations and after stroke. In their mouse model, the balance of the positive and negative effects of the complement system favoured an overall reduction of stroke damage.

BIOREMEDIATION

Cleaning up the Antarctic

Environ. Sci. Technol. doi:10.1021/es051818t (2006)
A five-year field study in the Antarctic has shown that fertilizing polar soils could help in the clear up of fuel spills.

Some estimates suggest that a million cubic metres of Antarctic soil has been contaminated with hydrocarbons, which leaked from vehicles and storage tanks. Shane Powell of the University of Tasmania in Australia and her colleagues investigated how organisms that live in the soil could help to remediate this damage. They identified some anaerobic denitrifying bacteria, which metabolize hydrocarbons, as being suitable for the job. But denitrifiers have been little studied in polar soils. Adding fertilizer pellets to the soil, which provided the denitrifying bacteria with nutrients, boosted the rate that petroleum hydrocarbons were degraded.

LIFE HISTORY

Double the fun

Curr. Biol. **16**, R117–R188 (2006)
Having sex doubles the lifespan of the lucky Ansell's mole rat, concludes a long-running study of *Cryptomys anselli* (pictured below).

Philip Dammann and Hynek Burda, both at the University of Duisburg-Essen, Germany, measured the lifespan of mole rats in captivity. Although the struggle to reproduce generally shortens lifespans in mammals, some breeding mole rats lived for 20 years, whereas all non-breeders were dead by their eighth birthday. The authors do not



H. BURDA

propose a mechanism for the effect, nor do they show whether the pattern exists in the wild where, like many insects, mole rats live in social colonies featuring a single breeding pair. But they do rule out the possibilities that breeding pairs enjoy higher status, better health or preferential access to food.

QUANTUM COMPUTING

The logical path

Science **311**, 1133–1135 (2006)
Even if quantum computers could easily be made, they wouldn't yet be able to do much; the algorithms they would need to run are hard to write and in short supply. Michael Nielsen and his co-workers at the University of Queensland in Australia claim that this deficit could be made up through the application of geometry.

In their analysis, the steps from logic gate to logic gate that make up a computation are treated as a path through the computer's 'space' of possible quantum states. The optimal algorithm will correspond to a geodesic in this space — the shortest path from input to output.

Such geodesics can be found using the trusty tools of Riemannian geometry, and this could mark a significant improvement on today's hit-or-miss approaches.

JOURNAL CLUB

Daan Frenkel
FOM Institute for Atomic and Molecular Physics, Amsterdam, The Netherlands

What to do when all the seats are taken? A materials scientist considers his options.

Adding particles to a crystal can be a bit like seating passengers in an aeroplane. As people board a plane, row after row of seats fills up until, at some point, they are all taken and the aeroplane is full.

In a crystal of fixed volume, there is a limited number of lattice sites to be filled by its passengers, such as atoms or molecules.

To seat more people without using a larger aeroplane, you must move the seats closer together. Likewise, I was taught — and I accepted — that moving the lattice sites of a crystal closer together was the only way to squeeze more particles into a solid of a given volume. This is a strategy resisted by aeroplane passengers, atoms and molecules alike.

However, computer simulations and theory presented in a recent paper (B. M. Mladek *et al. Phys. Rev. Lett.* **96**, 045701; 2006) suggest that another solution is possible: sufficiently soft particles should be able to form a 'cluster solid' with several particles per lattice site. This leads to the question, do these kinds of solids exist in nature?

The authors suggest that certain polymeric macromolecules could form such crystals, but experiments are lacking.

Interestingly, a real example of a cluster solid may come from a different field of physics, namely the study of quantum phase transitions (M.O. Goerbig *et al. Phys. Rev. B* **69**, 115327; 2004). This paper explains the observed behaviour of a thin layer of cold electrons in a magnetic field by assuming that the electrons form a crystal with more than one particle per lattice site.

This arrangement may be comfortable for particles, but the aeroplane equivalent is too horrible to contemplate.

NEWS

Disease surveillance needs a revolution

With avian flu spreading around the world at a frightening rate, scientists are welcoming an international proposal for state-of-the-art labs to monitor emerging diseases in developing countries. But they add that the bird-flu crisis has exposed glaring deficiencies that demand a radical rethink of the world's veterinary and disease-surveillance systems.

Avian flu is now endemic across large parts of Asia, and in the past few weeks has exploded across Europe and into Africa. "H5N1 has focused the spotlight of the world on disease surveillance, and it's showing up all the pimples and warts," says Bill Davenhall, who develops health mapping schemes for countries and is head of health at ESRI, a geographic information systems company in Redlands, California.

Developing countries, in particular, lack decent human-disease surveillance, and animal monitoring is often virtually nonexistent, with few basic laboratory and epidemiological resources available. "On the ground in Indonesia, there is no systematic programme at all," says Peter Roeder, a field consultant with the United Nations' Food and Agriculture Organization (FAO). "It's just a bloody mess."

Global danger

It is a problem that the developed world cannot ignore, because a disease that emerges in Bangkok or Jakarta could ultimately trigger a global disaster. So researchers at the US Department of Defense have suggested setting up a network of high-tech labs in developing countries to monitor cases of infectious disease (see page 25). The labs would be modelled on the US network of infectious-disease labs, such as the naval research unit NAMRU-2 in Jakarta. But they would be funded by the international community to avoid the local distrust that has often hampered labs run by the US military.

Such a network could vastly speed up and

"H5N1 has put the spotlight on disease surveillance, and it's showing up all the pimples and warts."

improve the diagnosis of viruses such as H5N1 when outbreaks occur, says Roeder. He points out that misdiagnosis of H5N1 as Newcastle disease in recent outbreaks in Nigeria and India led to long delays in control measures.

Mark Savey, an epidemiologist who heads animal health at France's food-safety agency, also welcomes the proposal, but cautions against the "mirage of technology" in surveillance. "You don't need satellites, PCR and geographic information systems to fight outbreaks," he says. The labs' top priority should be building large teams of local staff, who are familiar with the region and its practices, he argues. "If you do not have that, then surveillance will stay in the Middle Ages."

Savey recalls his trip to Russia last summer as part of a European team investigating outbreaks of avian flu. "You have a paper Michelin map; you have people who speak the language; you put red circles on outbreaks; and you use a pen and paper to compare them with things like the dates of market openings, and with how outbreaks line up with railways." Such local knowledge is crucial to interpreting data, he says. "If you don't know what the Trans-Siberian Express is like, with people cooped up for days, exchanging chickens and eggs at every stop, you would never guess that it was the Trans-Siberian that mainly spread avian flu across Russia."

Roeder agrees that the focus must be local. "No amount of setting international guidelines and publishing global action plans is going to help when you have an organization within the country that doesn't know what to do," he says.

Back to basics

But many feel that alongside setting up local centres, epidemiology needs a fundamental overhaul. Even in developed countries, the field has been chronically underfunded, says Antoine Flahault, director of Sentinelles, France's national disease surveillance network. He adds that he is jealous of the multimillion-dollar satellites that climate scientists enjoy and the powerful accelerators being built for physicists. In comparison, epidemiologists "are still in the nineteenth century," he says.

Flahault points out that his own national flu-monitoring system relies on a few hundred volunteer doctors submitting patient data to an online database. From these data, Flahault's team tries to build up a picture of disease across the country, spotting outbreaks and predicting when the year's flu season will hit.

The 122-city programme run by the US Centers for Disease Control and Prevention is the only one in the world where disease reports are made in real time, Flahault points out. And such lack of data prevents the field from developing sophisticated models of communicable disease. "It's as if we were trying to study the weather, but collected data only



Poor track record: the efforts of teams to monitor bird flu suggest the world must invest in epidemiology.



when there was a heatwave or storm," he says.

Ward Hagemeijer, the bird-flu programme manager at Wetlands International in Wageningen, the Netherlands, also complains of the general lack of resources. He has been on recent missions to sample H5N1 in affected

countries, but says he has been unable to get his African samples sequenced because certified labs have been too busy analysing samples from European outbreaks.

Leading light

Another fundamental problem is the lack of strong international leadership: there is no global body able to take overall responsibility for emerging diseases, particularly those that jump to humans from animals. The World Health Organization would be an obvious choice, but although it has a strong remit for public health, it is not responsible for monitoring outbreaks in animals — that duty belongs to the FAO and to the Paris-based World Organisation for Animal Health (OIE).

Neither of those organizations traditionally monitors outbreaks from a public-health point of view, however — the FAO is concerned with food safety and the OIE is responsible for trade issues. "Veterinary services throughout the world, particularly in developing countries, are very weak on this. They are not set up to watch for emerging disease events," says Roeder.

As well as a lack of expertise, the FAO and OIE do not have the funds for disease surveillance. "We need much more day-to-day interaction with locals on the ground, but we

haven't had the resources," Roeder admits.

The situation is slowly improving, however. The FAO and OIE had set a target of \$102.5 million for fighting flu in Asia, and by the end of 2005, countries had donated just \$25 million. But, in January, donors at a Beijing conference pledged \$1 billion in grants, and \$900 million in loans to support the FAO/OIE Global Strategy for the Progressive Control of Highly Pathogenic Influenza.

Many hope that resources pumped into avian flu will benefit the surveillance of emerging diseases generally by strengthening

ing infrastructure. But the money available so far is only a start. And the rate at which new diseases appear — currently around one a year — is increasing. Mark Woolhouse, an epidemiologist at the University of Edinburgh, UK, has

analysed all recognized human pathogens and he suggests that this rise is mainly due to changes in land use and the way that people live. Of the 117 emergent human diseases that he has studied, more than half jumped from animals.

"The world has to get to grips with the fact that what is happening now is going to happen repeatedly," says Roeder. "We have to develop a global structure to tackle emerging diseases. ■ Declan Butler

"Epidemiology has been chronically underfunded ... we are still in the nineteenth century."

Physicists told to confront those big questions

Time travel, multiple universes and extraterrestrial intelligence might seem more the purview of *Star Trek* scriptwriters than of serious researchers. But the scientists behind a new institute have announced their intention to change that perception.

On 27 February, the Foundational Questions Institute (FQI) made its inaugural call for proposals from scientists interested in asking the really big, and really odd, questions about the Universe — questions such as: why does time flow in a single direction, or, can intelligence survive in our Universe in the very long term?

"These are the very questions that a lot of scientists got into physics and cosmology to tackle," says Anthony Aguirre, a cosmologist at the University of California, Santa Cruz, and the FQI's

associate scientific director. "But they don't tackle them, because they either don't have time or don't have monetary support."

The institute's directors hope it will remedy this. The FQI was set up last October with a grant from the John Templeton Foundation, which

promotes research at the boundary of religion and science. With US\$8 million in seed money from the foundation, the FQI will fund dozens of researchers' part-time work on these questions, Aguirre says. All proposals will be peer reviewed, he adds.



Boldly go:
a new institute will fund work at the frontiers of science.

"I'm very happy to see that a project has started to address these needs," says Lee Smolin of the Perimeter Institute for Theoretical Physics in Waterloo, Ontario, who is also on the FQI's scientific advisory board. Smolin says he believes the project will help shake up the current culture, "which emphasizes technical brilliance over ambition and originality".

But not everyone welcomes the institute's unusual remit. There is no shortage of crazy ideas in theoretical physics, says Paul Steinhardt, a cosmologist at Princeton University, New Jersey. "Metaphysics is running rampant through string theory and cosmology," he says. "I would like to see things go a little bit in the other direction."

Geoff Brumfiel
♦ www.fqxi.org

SCIENCE IN THE MOVIES

Check out the news site after this weekend's Oscars for a celebration of science in film.

www.nature.com/news



AMIPAS

R. VOGEL/AP PHOTO

PARAMOUNT TELEVISION/THE KOBAL COLLECTION

Bets are placed for next Harvard president

The repercussions of Larry Summers's resignation as Harvard president are being felt both within and beyond the university's courtyards. His exit has cast a shadow on plans to promote science at the institution, and highlights the challenges faced by university leaders worldwide.

Summers's announcement on 21 February that he will step down in June marked the climax of a year of intense criticism from academics unhappy with his brusque leadership and lack of consultation with faculty. Former Harvard president Derek Bok will stand in as interim leader until a new president is chosen.

Summers's departure leaves some Harvard scientists in a state of disquiet. One of his central ambitions for the university was to develop the sciences, particularly interdisciplinary research, which he viewed as essential for Harvard to remain at the forefront of academia (see *Nature* 433, 190–192; 2005). "There is a sense of unease," says David Scadden, co-director of the Harvard Stem Cell Institute, one of the projects championed by Summers. "We want the successor to share his sense of priority."

One of Summers's boldest plans was to place science facilities at the core of a vast new

campus to be built in Allston, Boston. Just days before his resignation, the university unveiled concrete plans for the first building on the Allston campus, a 46,450-square-metre complex that will house the Stem Cell Institute among other science initiatives.

Steven Hyman, Harvard provost and top administrator, told *Nature* that expansion of science remains a main priority for the Harvard Corporation, the governing body responsible for selecting Summers's successor. "The university remains absolutely committed to a vision of building science at Harvard," he says. But because most details have yet to be thrashed out, some academics say Summers's departure provides an opportunity to reassess the plans, for example, which areas of research merit investment. "There are concerns about

"The university remains absolutely committed to a vision of building science at Harvard."

the direction to go and how fast to go," says physicist Gary Feldman, one of Summers's critics. "We need a careful and impartial assessment of Harvard's needs."

With so much at stake, all eyes are on the selection process for the next president. Many faculty chafed at Summers's corporate management tactics — he previously served as US treasury secretary. The selection committee



Who next? Harvard scientists are hoping the next university president continues to prioritize science.

HARVARD UNIV.

will therefore probably choose someone already senior at another university or from within Harvard, predicts William Funk of executive recruitment firm Korn/Ferry International, who helps universities find new presidents. "I think this time they'll be very risk averse and choose someone steeped in academe with a proven track record," he says.

Because of Harvard's position at the pinnacle of the US university system, Summers's resignation is attracting intense scrutiny from academics nationwide. They say Summers's

Scepticism greets EU plan for virtual institute

MUNICH

Plans announced last week to create a European Institute of Technology (EIT) have received a lukewarm reception from the scientific community.

The new body is being touted as the European counterpart of the US Massachusetts Institute of Technology (MIT). It will aim to translate research into commercial opportunities and jobs. But many are sceptical about what the 'virtual' institution, which would consist of partnerships between academic and industry research groups across Europe, is likely to achieve.

José Manuel Barroso, president of the European Commission, outlined plans for the EIT on 22 February. He said it would be "a light and flexible organization" that will foster education, research and innovation in commercially promising fields of science. But he gave few details about how much the institute is likely to cost, or the research areas it will target.

A proposal to create a centre for the EIT in the European parliament building in Strasbourg was discarded, mainly because choosing a single site would have been interpreted as favouring a particular

country or region. So the institute will be made up of partnerships in which members of research groups working on a particular area will spend time in each others' labs. If approved by the European Council later this month, the initiative could be up and running by 2010.

But scientists doubt that the proposed network will yield anything as efficient as MIT, with its annual budget of \$2 billion. The proposed structure is scarcely different from research collaborations already funded by the European Union, says Renzo Rubele, a theoretical physicist at the University of Salerno in Italy,

and a member of an expert group on university research. To make a real difference, he adds, would require a powerhouse like the particle-physics lab CERN in Geneva, Switzerland, where people and ideas are continuously exchanged.

Science managers are also unconvinced. "In principle, a European MIT would be a good thing. But at this stage it is an ill-prepared political rush job," says Helga Nowotny, vice-president of the scientific council of the proposed European Research Council and former chair of the European Research Advisory Board.



Rules tighten for stem-cell studies

Researchers have taken a first stab at clarifying the conflicting rules that govern stem-cell research. A panel of more than 50 scientists, journal editors and ethicists have produced guidelines that they hope will promote international collaborations in a field badly shaken by the South Korean cloning scandal.

In Germany, most stem-cell research is illegal, and researchers could in theory be jailed if they participate in embryo experiments abroad. China lacks clear national policies, with different institutes following different rules. Laws also differ between US states. "The situation is likely to get worse," adds Robin Lovell-Badge, a stem-cell biologist at the National Institute for Medical Research, London, and member of the panel. He says even countries with well-developed laws must continually update them to cope with emerging techniques.

The panel will set up a website where researchers can explain local laws and share existing ethical guidelines. It also wants journal editors to ask authors for more information relating to the ethics of the research they publish, such as the source of cells, to encourage the development of transparent ethical standards.

The group stresses the importance of obeying national laws, but calls on countries such as Germany to reconsider legislation that extends to research elsewhere. "Laws all around the world are stopping international collaboration," says panel member Julian Savulescu, an ethicist at the University of Oxford, UK.

The meeting, which took place over 22–24 February in Hinxton, UK, took two years to plan. It was not designed to tackle fraud, such as the case of Woo Suk Hwang, the South Korean scientist who faked data showing stem cells could be extracted from cloned human embryos. But the group does propose that egg donors are treated in the same way as human research subjects, meaning that Hwang's use of eggs from his researchers would probably not have been approved.

The group plans to meet again, probably within two years, to discuss new technologies such as chimaeras. ■

Jim Giles

For the Hinxton Group's statement

► www.nature.com/news/2006/060227/full/440009a

run-ins with faculty reveal the difficulties faced by a modern university leader, who must increasingly spend time away from campus raising funds and liaising with the wider community.

Although it is easy to develop sweeping plans, a good director must be able to sell them to faculty and staff, says Ingrid Moses, chancellor of the University of Canberra, Australia, and former president of the International Association of University Presidents. "They are the ones who are crucial for implementation," she says, "hence the importance of interpersonal skills." ■

Helen Pearson

"What Europe really lacks in terms of innovation is sufficient venture capital for spin-off companies," she says. "The EIT cannot solve that problem."

Nowotny adds that although clusters of academic and industrial research work well at the regional level, she doesn't believe sending industry researchers to distant university labs will boost innovation. Barroso is confident that industry will take the bait, however. This could be crucial for the EIT, as it is still unclear how the institute will be financed.

Several European companies, including Pirelli and Nokia, have welcomed the idea, but have not specified what they might contribute. Help may also be on the way from across the Atlantic: Microsoft founder Bill Gates has signalled that his foundation intends to support the research flagship. ■

Quirin Schiermeier

Demo backs animal lab in Oxford

OXFORD

Hundreds of scientists and students took to the streets of Oxford last weekend to march in favour of the city's new biomedical lab. The rally was the first time that proponents of animal testing have come out en masse to support their cause. It happened in response to threats from animal liberationists who claim that anyone associated with the University of Oxford is a legitimate target for physical violence.

The building of the Oxford lab has been hampered by persistent threats from animal-rights extremists. The project has been the target of regular protests by antivivisection campaigners, but previously only a handful of researchers had raised their voices to support the lab — most of its backers, including the university itself, had remained silent on the issue.

The event, held on 25 February and attended by around 600 marchers, represents a change of strategy amid swelling grassroots support for the lab and growing indignation over the intimidating tactics of animal-rights extremists. "The university was not keen on



Battle lines: Saturday's march was the first to support the building of an animal-research lab in Oxford.

raising the ante with a public protest," says Evan Harris, a local member of parliament. "But once there was this threat to Oxford students, I think the balance changed."

Harris describes the event as "the best kind of rabble-rousing", with protestors chanting slogans such as: "No more threats, no more

fear; animal testing's wanted here."

"This is an historic day — we're drawing a line in the sand," said Oxford neuroscientist John Stein as he addressed the rally. "We must not be intimidated. It's not just about research; it's about democracy."

Support for the event was fomented by

Sackings put editorial freedom on the spot

The abrupt sacking of two editors of a medical journal has been blamed on political differences between them and the journal's owners. With the editorial board fighting to have the two reinstated, the case has focused attention on the issue of editorial freedom for journals owned by academic societies.

John Hoey, former editor of the *Canadian Medical Association Journal* (CMAJ), and his deputy Anne Marie Todkill were sacked on 20 February. Since then, 16 of 18 editorial board members have signed a letter asking the CMA to reinstate them. But Graham Morris,

president of CMA Media, the company that publishes the journal, denies his decision had anything to do with editorial freedom.

"Hoey did a good job in building the reputation of the journal," says Morris, adding that he is simply looking for "a fresh approach".

But Hoey and Morris have often locked horns. For example, Hoey ran an editorial in the 3 January 2006 issue detailing changes he says he was forced to make to an article about emergency contraception.

"I think it is extremely damaging for the reputation of the journal."

As part of a news story on whether pharmacists were asking users unnecessary questions before dispensing the 'Plan B' pill, the

CMAJ asked women across Canada to buy the pill from local pharmacies and report their experiences. According to Hoey, the Canadian Pharmacists' Association pressured

the CMA into insisting that the investigation could not be published without peer review. Morris says he wasn't influenced by the pharmacists, and that the

decision to remove the section about the investigation left a "well-balanced" piece.

There were also deep-seated differences between the journal editors and the CMA over privatization of health care, according to Trudo Lemmens, a medical ethicist at the University of Toronto. In August, the CMA voted to support a Supreme Court ruling that ended the ban on private health care in Quebec. The CMAJ however, is perceived to be against any erosion of the nation's public medical system. According to Hoey, a news story on incoming health

Laurie Pycroft, a 16-year-old blogger who was moved by what he saw as the one-sidedness of the animal-research debate. Until now, public demonstrations have been the preserve of groups opposed to animal research; scientists have mostly kept their heads below the parapet, for fear of personal reprisals. Pycroft created a research-advocacy website called Pro-Test, under whose banner the rally took place. "I'm extremely proud of the turnout," he told *Nature*. "I would have been happy with 50 people."

Meanwhile, about 150 antivivisection campaigners mounted an opposing demonstration. This event, culminating in a rally at the exact spot where pro-research campaigners had mustered just hours earlier, was part of a series of demonstrations by the animal-rights group Speak.

Demonstrators at the Speak rally noted the emergence, for the first time, of large-scale public opposition to their views. One demonstrator called for the issue to receive even wider coverage now that battle lines have been formally drawn. "We want to see a fair and unedited debate on prime-time television," he said.

It is a debate that traditionally has been played out in actions rather than words. The Oxford lab recently endured an 18-month hiatus in construction after the building company Walter Lilley pulled out in response to



Laurie Pycroft (right) founded Pro-Test and arranged the rally.

S. HIRD/REUTERS

"We must not be intimidated. It's not just about research; it's about democracy."

bad publicity and diving share prices. Other research facilities, such as Huntingdon Life Sciences near Cambridge, have also been subjected to intimidating tactics and violence from extreme animal-rights groups.

Speaking at the Pro-Test event, Simon Festing, director of the London-based Research Defence Society, called animal research a "moral choice" and "a price worth paying." And Oxford neuroscientist Tipu Aziz said the march was aimed at "defending our right to better humanity's plight."

For others, the event was about standing up against what is perceived as bullying from extremists. "We're sick of the intimidation," says postgraduate research student Kristina Cook. "It's amazing how many people support what we do but are afraid."

Michael Hopkin

minister Tony Clemens, which included critical quotes about Clemens's acceptance of privatization was pulled by the management.

Jerome Kassirer, a member of the *CMAJ's* editorial board, points out that the journal has been doing well, making Morris's call for a change seem oddly timed: "Given that the journal was getting better, was more respected, and its impact factor was going up, it's hard to escape the conclusion that they were fired for political reasons."

Fiona Godlee, editor of the *British Medical Journal*, says she was "shocked and saddened" by the news. She

worries that the *CMAJ* may now recede into relative obscurity. "I think it is extremely damaging for the reputation of the journal."

Godlee hopes the firings will stiffen the resolve of journal editors to fight for a strong wall between their editorial content and the politics of home associations. "It is beholden on editors in general to hold the line for the sake of maintaining independent scientific voices around the world."

Also "very concerned" about the sackings is Shabbir Alibhai, a specialist in cancer and geriatrics at the University of Toronto and member of the

CMAJ's reader panel. But Alibhai says the episode raises questions about the role of journal editors, particularly when writing editorials. "Who is the editor representing when he or she pens words?" he asks. "The readers? The association? The journal? Or just the editors?"

Meanwhile editorial board member Philip Devereaux is widely distributing a petition to reinstate the editors. "This is one of the most fundamental battles, and it must be fought," he says. "It's not just because John and Anne Marie are really nice people." Emma Marris

ON THE RECORD

"Who would have thought you'd have standing room only at a geek event?"

Immunologist John Cohen is astonished at how Café Scientifique — a series of informal debates with scientists — has taken off in Colorado.

Source: *New York Times*

SCORECARD



Dog turds

A California company announces plans to use a 'methane digester' to extract energy from faeces scooped up at a popular San Francisco park.



Thermal underwear

Employees at Japan's environment ministry bundle up to stay warm, after turning off the heat to help meet the country's targets for greenhouse-gas emissions.



Blackbeard

Engineers are building a sand dune underwater to protect what is believed to be the famed pirate's ship, sunk in 1718 off the coast of North Carolina.

DATAPOINT

How much do people in different countries care about science? A glimpse comes from the Science and Engineering Indicators 2006, published by the US National Science Board. People indicated their interest in new scientific discoveries on a scale from 1 to 100.

China	70
United States	69
Europe	54
South Korea	50
Japan	44

NUMBER CRUNCH

Earth crossed another population milestone last weekend.

6.5 billion is the world's population, as of 25 February.

131 million people will be born in 2006.

57 million people will die in 2006.

Source: *US Census Bureau*

Alternative energy plan criticized

WASHINGTON DC

Just weeks after President George W. Bush vowed to wean the United States off oil and onto alternative energy sources, scientists and policy analysts are voicing concern that the research push is unlikely to yield the promised benefits.

Their worries were highlighted last week by events at the National Renewable Energy Laboratory (NREL) in Golden, Colorado, the Department of Energy's main research centre for energy sources such as solar, wind and ethanol. Just days after being laid off, 32 NREL staff were reinstated when it emerged that Bush was going to visit the lab during a tour to promote his focus on alternative energy.

During his tour of the lab on 21 February, Bush blamed poor communication for the oversight. "I recognize that there have been mixed signals when it comes to funding," Bush said. "Unfortunately, there are sometimes decisions made, but as a result of the appropriations process, the money may not end up where it was supposed to have gone."

Bush's 'advanced energy initiative' made its debut in his State of the Union address on 31 January, when he called for a 22% funding increase for research into alternative energy technology.

But budget analysts looking at the president's 2007 budget request say the proposed increase is more cautious, with funds moved into some research areas and removed from others. For instance, there is more money for research on solar energy, with a rise of 79% to \$148 million, and biomass, up 65% to \$150 million. But funding for energy conservation is down 6.3% to \$289 million, and the geothermal programme is axed altogether.

"The wider your view, the less glamorous it looks," says Kei Koizumi, a budget analyst at the American Association for the Advancement of Science. "If you look just at biomass,



George W. Bush (right) toured the National Renewable Energy Laboratory to promote his policy on alternative energy sources such as wind power.

it looks great. If you look at all renewables, it looks less great."

Energy research, including work on fossil fuels, would

decline overall under the president's budget, says Koizumi. He adds that there is no money set aside for later years for the energy initiative.

Reaction elsewhere has been mixed. Solar enthusiasts are pleased with a budget increase for work on photovoltaic cells. Noah Kaye, spokesman for the Solar Energy Industries Association, called it "a key victory for a growing high-tech industry in the United States", but went on to call for production incentives such as tax breaks.

Proponents of wind energy sang the same tune, only with less enthusiasm. Their research boost is just 13% to \$44 million.

The president's call for increased funding got a mixed reception from environmentalists. The funds just aren't enough, they say, and are too focused on research. "We need other policies including technology incentives, or caps on emissions," says Andrew Aulisi, a senior associate at the Washington-based World Resources Institute. "You need lots of different policies to get a handle on the climate and

energy crisis. Even within R&D, the numbers are not that good."

To make matters worse, analysts point out that a large part of the increased money is likely to be taken up by earmarks, in which legislators appoint money to projects in their home states. For example, the NREL has blamed the lay-offs on the large number of earmarks in the 2006 budget, which it says left it with a \$28-million deficit in operating costs.

Earmarks are rare at agencies such as the National Institutes of Health and the National Science Foundation. But they made up 21% of the energy research and development budget last year, which is the highest ever, according to Koizumi.

"The energy department has earmarks that they have no choice but to fund," adds George Douglas, a spokesman for the NREL. "It is taking away money that could be used in this type of research."

Most scientists would favour peer review as a way of distributing the energy department's research funds, says Marchant Wentworth, legislative representative for clean energy at the Union of Concerned Scientists. "In the short term it does no good to have an on-again, off-again programme."

Jacqueline Ruttiman and Emma Marris

Crashed polar satellite earns reincarnation

An Earth-observation satellite that failed just minutes after its launch last October is to be replaced, the European Space Agency (ESA) announced on 24 February.

The €140-million (US\$165-million) CryoSat, which was designed to monitor Earth's shrinking ice caps, fell into the Arctic Ocean north of Greenland when the second stage of its launch rocket malfunctioned (see *Nature* 437, 1078; 2005). Mission scientists subsequently lobbied ESA and pushed for a rebuild, a request that has now been granted. The space agency says that a second CryoSat can be built for €106 million using infrastructure already in place for other missions.

CryoSat's second incarnation should be ready for launch in 2009.

Argentine fossils seized at Arizona gem show

US authorities have raided a major gem and mineral show in Arizona, marking the first crackdown on the widespread sale of international fossils there. Such sales have long been the scourge of palaeontologists.

On 10 February, federal agents seized about 6 tonnes of Argentinian fossils, including at least three dinosaur eggs valued at more than US\$4,000 each. For two weeks each winter, hotels and vacant parking lots in Tucson are turned into geological marketplaces for the show (see *Nature* 403, 690; 2000).

Acting on a complaint from the international police organization Interpol, agents confiscated the fossils, which were being sold by the Buenos Aires-based firm Rhodo Co. Rhodo officials could not be reached for comment.

A criminal investigation has been opened

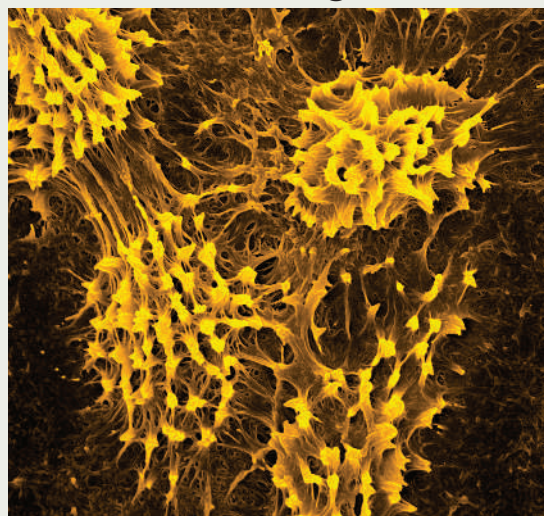


Precious cache: illicit fossils, including dinosaur eggs, were found among minerals at a US show.

Prizewinning stem cells take Australian grand tour

Both beauty and scientific merit are lauded in this image — the winner of an Australian photography competition focused on stem cells. The photo shows human embryonic stem cells developing into epithelial precursor cells.

A collection of 35 photos, intended to engage public interest in stem-cell research, is being displayed around the country. Currently in Sydney, the exhibit has been on show in parliament buildings, business offices, hospitals and universities. The competition was organized by the Australian Stem Cell Centre in Melbourne.



L. STAMP & J. CLARK

to determine whether US or Argentinian laws were violated when the fossils were shipped to Tucson for sale.

'Shadow' debate on climate looks beyond Kyoto

Politicians, business and development leaders have launched their own dialogue for tackling climate change after the initial phase of the Kyoto Protocol ends in 2012.

Frustrated by the slow pace of government action, the group plans to shadow the formal United Nations negotiations on a successor to the Kyoto agreement, which limits greenhouse-gas emissions. They will explore ways in which developing countries can cooperate to reduce emissions.

The group includes lawmakers from the G8 industrial nations, as well as India, China, Brazil, Mexico, Spain, Australia and South Africa. No representative of the US Congress was present at the 24 February launch, although the effort's organizers say they hope to have significant US participation.

Margaret Beckett, the UK environment minister, said she hoped the discussions would "encourage more free thinking" than formal negotiations. The group's work will culminate in a report to be presented at the Japanese G8 summit in 2008.

Fake data did not harm nuclear-waste study result

The inclusion of false data in documents relating to a proposed nuclear-waste repository at Yucca Mountain, Nevada, has not altered conclusions about the site's scientific worthiness, says a report from the

US Department of Energy. But the department recommends that much of the work should be redone anyway.

The 17 February report comes almost a year after revelations that scientists at the US Geological Survey faked the times and dates of geological samples used to study how water seeped through the mountain (see *Nature* 434, 427; 2005). Such data are crucial because they indicate how long it might take for nuclear waste to enter the water supply if the containers were to leak.

Independent data support the water-seepage estimates, the report concludes.

Licensing thumbs-down for drug from GM goats

The European Medicines Agency last week turned down a company's request to market a drug made in the milk of genetically modified (GM) goats. The decision means that, despite more than a decade of using GM animals to produce drugs, no products have yet been approved for use.

GTC Biotherapeutics of Framingham, Massachusetts, developed a herd of GM goats that produce milk containing a human anticoagulant called antithrombin. The company planned to market the drug, under the name ATryn, for people with a genetic defect that makes their blood prone to clotting.

But the medicines agency turned down the request on 23 February, saying the product had not been tested thoroughly enough.

Attention now turns to a Dutch company, Leiden-based Pharming, which has an application pending with the US Food and Drug Administration to market an antibacterial agent produced in the milk of GM cows.

FOCUS ON THE LIVING

Atomic force microscopes have revolutionized the study of materials, but probing watery biological systems has proved more difficult. **Jenny Hogan** asks whether a fix is at hand.

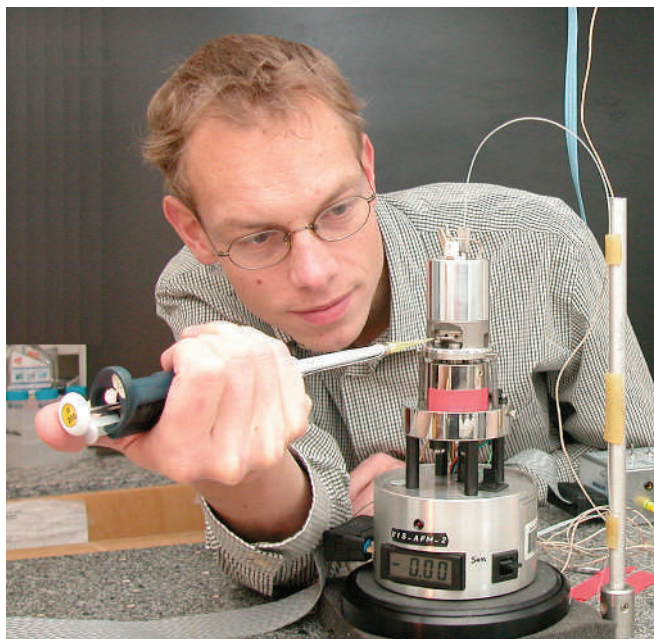
Bart Hoogenboom's windowless room is a cramped space, almost filled by the sturdy table at its centre. Cables dangle from piles of electrical devices that climb to the ceiling. In the midst of the tangle of equipment is a stack of three small metal cylinders. Hidden inside the top one is a sliver of silicon, its tip quivering up and down 200,000 times each second.

"It is an experimental physicist's dream and a biologist's nightmare," says Hoogenboom. Happily, Hoogenboom is a physicist, a post-doc at the University of Basel in Switzerland, and for him the roomful of gadgetry to tinker with is a treat. But the outcome of his tinkering could enrich the lives of biologists.

The cylinders on the table are the working parts of an atomic force microscope, or AFM. Invented 20 years ago¹, AFMs are based on a sharp tip at the end of a flexible beam, or cantilever; the topography of a surface is detected by the bending of the cantilever as the tip scans the specimen. Under the right conditions, AFMs can produce images that are accurate down to the last atom.

The AFM has revolutionized the way material scientists study surfaces. But in general, it has been harder to apply AFMs to the study of delicate biological molecules and the soft tissues in which they are found. The team Hoogenboom works in, and others in labs around the world, are planning to change that, by finding out how to use AFMs in ways that don't damage the samples.

Biologists already have pretty good methods for studying things down to molecular or even atomic scales, but these involve taking the objects of study out of their normal biological context. Electron microscopy requires samples to be fixed and exposed to a vacuum; X-ray crystallography requires the relevant proteins to be forced into crystalline arrays. "The advantage of the AFM is clear," says Peter Hinterdorfer, a biophysicist at the University of Linz in Austria, who has used the technique to



Physicist Bart Hoogenboom's love of gadgetry could pay off for biologists.

study how antibodies bind to their targets: "you can image in physiological conditions."

Some of the most impressive biological AFM images, including the image shown opposite, are the work of Andreas Engel, the head of Hoogenboom's team. But even this beautiful picture highlights the limitations of the technique. The golden rings are rotors that form part of an energy-conversion machine found in cell membranes, called an ATPase. Each protein ring has a diameter of around 5.4 nm — and although it is possible to see the molecules that make up each ring, you can't make out the amino acids of the protein, never mind see the thousands of atoms from which the rings are built.

Close encounters

This is poor resolution compared with that possible for hard surfaces, but is good for biological images taken with AFMs in 'contact mode'. AFMs can work in a bewildering variety of ways, but contact mode is the most straightforward. The sample is brought into contact with

the tip and moved back and forth and from side to side. But it is not the most precise way of doing things. The AFM tip tends to damage or dislodge the things it is scanning, which degrades the image resolution and limits the types of sample the technique can be used to study.

The obvious solution is to make the tip behave more gently when scanning the samples, and it is to this end that various researchers have turned to the 'frequency modulation' or FM mode. In FM mode, the tip hovers just above the surface under study. A pulse of energy is used to make the cantilever tremble, and the topography of the surface below affects the frequency at which it does so. So long as the forces that the tip senses as it hovers above the surface can be inferred from changes to the cantilever's frequency, the tip doesn't actually have to touch the sample at all.

AFMs used in FM mode (FM-AFM) can achieve stunning results — even resolving features below the scale of an individual atom, thought to be the signature of atomic orbitals, on a hard, flat silicon surface². But that sort of resolution requires the system to be used in a vacuum. For a long time, getting the FM mode to work in a messier, wetter environment seemed out of the question.

Having water around the cantilever was expected to deaden the vibrations. "Most experts would not have expected FM-AFM to work in water," says Stuart Lindsay of Arizona State University in

Tempe, whose group develops AFMs and other scanning-probe instruments for biological applications. "I would have been included in that class. Otherwise, I would have tried it first."

The first evidence that FM-AFM might work in water came from Takeshi Fukuma of Kyoto University's electrical-engineering department in Japan. When Fukuma was a postdoc in Hirofumi Yamada's lab, which does research into molecular electronics, he sought to reduce the noise in the AFM's output by optimizing the

"Being able to see small molecules binding to proteins would vibrate the whole community."

— Daniel Müller

P. REIMANN

way the cantilever's deflection was being measured. He thought this might compensate for the water damping down the oscillations.

In most commercial AFMs, a laser beam is bounced off the top surface of the cantilever to work out the tip's position. Fukuma and his co-workers set to improving each element of this sensor system, boosting the laser power and introducing a technique used in CD-ROM and DVD drives to stabilize the laser performance. The combined effect of the changes was to reduce the noise in the system to near its theoretical limit³. Without even trying to minimize the noise from other sources, such as the buffeting of the cantilever by water, the team got striking results. "I did just a few experiments to get atomic resolution," says Fukuma. In July last year, the research team reported their findings⁴. The paper's centrepiece is an image of mica taken in pure water, showing the atoms which are just half a nanometre wide.

Scaling down

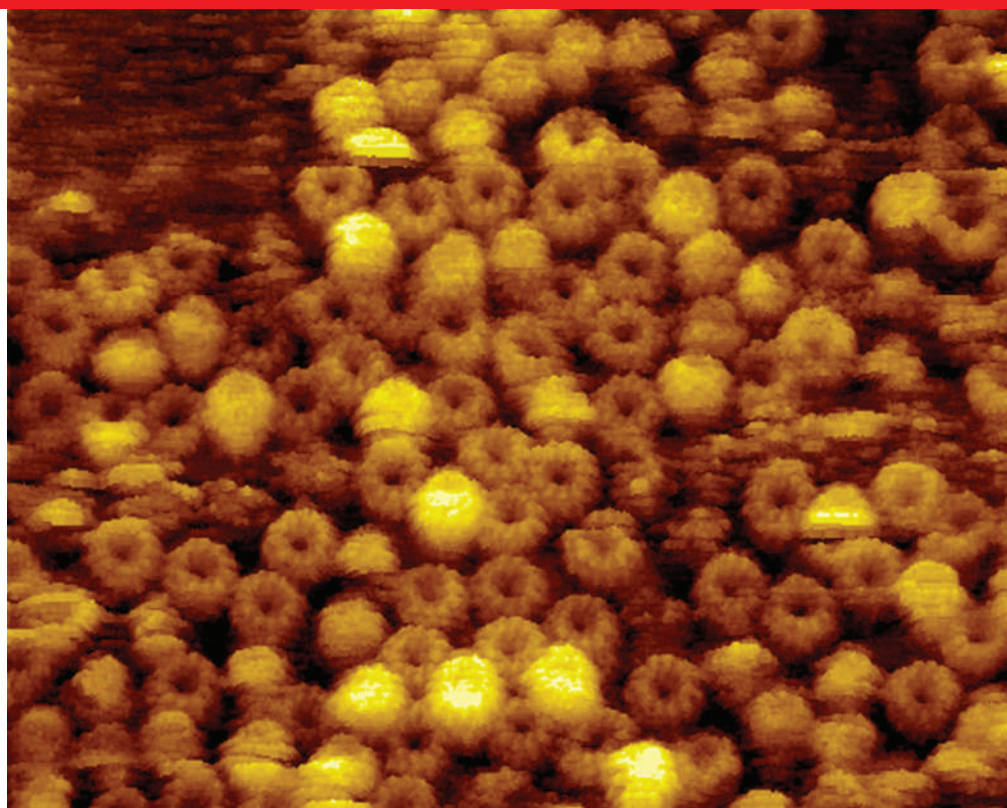
The developments in Yamada's lab are attracting attention. "This is certainly an innovation we're interested in and something we're watching," says Craig Prater, director of technology development for Veeco Metrology in Santa Barbara, California; Veeco is one of the world's leading suppliers of AFMs. Yamada says that Seiko Instruments and Jeol, two other companies that build AFMs, have also expressed interest in his work.

But the advance is not yet on the radar of the researchers whom it might benefit. "Biologists will pay attention as soon as the technique is used to image biological samples," says Daniel Müller, who collaborated with Engel on the ATPase work, and is now at the University of Technology in Dresden, Germany.

Yamada has recently presented images of biological samples, such as the membrane protein bacteriorhodopsin, at conferences. And looking for a closer collaboration with biologists, Yamada's student Fukuma has now moved to work with an interdisciplinary group at Trinity College Dublin in Ireland headed by Suzi Jarvis. Within three months of arriving in Jarvis's lab, Fukuma had optimized its AFM, as he had done in Kyoto, making many of the electrical components himself.

Fukuma's results have spurred Hoogenboom's group to ever greater efforts. His team has now achieved true atomic resolution of mica in water too. It worked independently of the Kyoto group on optimizing the measurement of cantilever deflection, completely replacing the optical beam deflection system used in standard AFMs with a component designed by the team⁵ — hence all the clutter.

The Basel team has also produced images



Atomic force microscopy has revealed the detailed structure of energy-converting ATPase rotor rings⁶.

of bacteriorhodopsin (a paper on this is under review at *Applied Physics Letters*). The FM-AFM images of bacteriorhodopsin seem slightly fuzzy compared with the best made in contact mode. But that doesn't worry Hoogenboom.

"We have ten years' experience doing contact mode; we haven't yet optimized the FM technique," says Hoogenboom. "If you drive a car for the first time, you don't know how fast you can take a corner."

FM-AFM is unlikely to achieve atomic resolution for biological samples, however. AFM is in general a rather slow technique, and during the minutes that a scan takes, the molecules under study will wriggle. Also, the sides of the tip will interact with any portion of the sample that rises up near it. Several groups are working on reducing these effects, for example by making the tip sharper or speeding up the process of scanning.

But even a small improvement in resolution over the contact-mode images could help, according to Müller, who worked on the ATPase rotors. Just being able to see small molecules binding to proteins "would vibrate the whole community," he says. This could provide, for example, a quick way to see how drugs bind to their target receptors — providing a new tool in drug discovery.

But until people start to use FM-AFM, no-one is really sure what the implications will be. "There are some interesting results coming out now," says Jason Cleveland of Asylum Research, an AFM supplier that is collaborating with Jarvis's group to explore how to introduce FM-AFM capability to their instruments. "But it's right at the beginning of the snowball, and you don't really know how big it will be."

Back in Basel, Hoogenboom recalls a discussion at a conference where biologists explained how they wanted an AFM to work. Their request was for a black box where you press a button and it gets you nice images. The FM-AFM technique is more complex than that, but the instrument needn't be a nightmare, assures Hoogenboom. "A bit more

development and it will almost look as simple as your CD player."

Jenny Hogan is a reporter for *Nature*.



Fix finder: Takeshi Fukuma boosted the efficiency of an atomic force microscope.

1. Binnig, G., Quate, C. F. & Gerber, Ch. *Phys. Rev. Lett.* **56**, 930–933 (1986).
2. Giessibl, F., Hembacher, S., Bielefeldt, H. & Mannhart, J. *Science* **289**, 422–425 (2000).
3. Fukuma, T., Kimura, K., Kobayashi, K., Matsushige, K. & Yamada, H. *Rev. Sci. Instr.* **76**, 126110 (2005).
4. Fukuma, T., Kobayashi, K., Matsushige, K. & Yamada, H. *Appl. Phys. Lett.* **87**, 034101 (2005).
5. Hoogenboom, B. W. et al. *Appl. Phys. Lett.* **86**, 074101 (2005).
6. Stahlberg, H. et al. *EMBO Rep.* **2**, 229–233 (2001).



Shaking the foundations

Last autumn's deadly earthquake caught Pakistan's government and scientific community off guard. Now a handful of officials and academics are struggling to bring the country up to code. **Geoff Brumfiel** reports from the scene.

On the morning of Sunday 8 October, residents of Pakistan's capital, Islamabad, were shaken from their beds. The city was in the grip of an earthquake so violent that people could barely stand to flee. On the western side of town, a poorly constructed apartment building collapsed, killing dozens.

Closer to the quake's epicentre, the story was worse. In the northern city of Muzaffarabad, concrete buildings pancaked, killing their residents instantly. Massive landslides wiped out villages perched on steep mountain slopes, and falling rock severed the narrow highway that connects the mountainous region to the rest of the country. Phone lines were down, but rumours were flying. Some said nuclear testing had caused the event.

The earthquake caught most Pakistanis off

guard, says Khawaja Azam Ali, dean of the faculty of natural sciences at Quaid-i-Azam University in Islamabad. The last time Pakistan experienced a massively destructive earthquake was in 1935, when the western city of Quetta was flattened, killing more than 30,000 people. This time, says Ali, "knowledge about the earthquake risk was zero. There were simply not many people thinking about it."

That's surprising, because winding through the foothills just beyond Ali's office window is a geological fault hundreds of kilometres long. It is one of about half a dozen similar faults running more or less through the heart of Islamabad. They are a silent reminder that this troubled nation — home to grinding poverty, tribal insurgencies, radical Islam, and more than two dozen nuclear warheads — lies in one of the world's most seismically active regions.

In the wake of last October's disaster, which killed at least 73,000 people, a fledgling movement has emerged to prepare the nation for future earthquakes. The government has pledged millions of dollars to build a new seismic network, and local universities are ramping up programmes in seismology and earthquake engineering. But success will require a sea change in Pakistan's attitude towards seismic risk.

The nation's military government will have to release sensitive data that have remained hidden. Universities must train a new community of seismologists. And ultimately, ordinary people — many of them illiterate — will have to be educated about how to survive quakes. This effort will take collaboration between the military, civilian government and educational institutions on a scale never before

TARIQ MAHMOOD/AFP/GETTY IMAGES

seen in the country. But it can be done, says Ali: "If there is such cooperation, then yes, there is a chance of progress."

The problem is formidable. Pakistan lies at the junction of three tectonic plates (see graphic). From the southeast, the Indian plate is sliding towards Afghanistan at some 40 millimetres each year, while from the southwest the Arabian plate moves northwards at a nearly identical clip. Caught in this geological pincer movement is a small promontory of the massive Eurasian plate. As the Indian and Arabian plates plough into Eurasia, they push up the Himalayan and Hindu Kush ranges in northern Pakistan, and a series of smaller ranges along its borders with Afghanistan and Iran. The enormous compression warps and tears the plates, creating hundreds of active faults throughout the country.

Nearly all of the major faults have been catalogued by the Geological Survey of Pakistan. But the survey did little beyond mapping each fault. To assess the danger, seismologists need a record of the tiny tremors that rattle the country almost daily. Collecting data from multiple seismic stations on those smaller quakes shows which faults are currently under stress.

Such 'microseismic' data have been collected for years by three public agencies: the Water and Power Development Authority, which oversees the nation's dams; the Atomic Energy Commission; and the Meteorological Department, which is formally responsible for seismic monitoring. But historically, that information has not been made public, according to



"The paranoia is completely unfounded. If you want to know exactly where something is, you can get a GPS receiver from Wal-Mart."

— Roger Bilham

Qamar-Uz-Zaman Chaudhry, director-general of the meteorological department. One main reason, he says, was the army's concern that the data could be used to monitor explosive tests related to the nation's nuclear programme in the lead-up to its 1998 nuclear tests. "Previously there were fears about nuclear monitoring, and that has hampered our cooperation with other countries," Chaudhry says.

Pakistan's army has another reason for secrecy: the country's most active seismic areas are in the contested region of Kashmir. Since the partition of India and Pakistan in 1947, both countries have laid claim to the region. To this day, roughly 100,000 Indian and Pakistani troops stare at each other across a delicate line of control. Fearing that either side could gain even a tiny edge from geospatial measurements, "both India and Pakistan have said that nobody can do any kind of GPS measurements", says Jack Schroder, a seismologist at the University of Nebraska in Omaha who has worked in Pakistan.

"This paranoia is completely unfounded," contends Roger Bilham, a geologist at the University of Colorado, Boulder. "Aerial photographs are available to everybody, maps were left by the British, and if you want to know exactly where something is, you can get a GPS receiver from Wal-Mart. If the government wants to save people, they've got to protect them from earthquakes, not India."

In fact, there are signs that Pakistan's military government is opening up. Immediately after the October quake, government officials asked non-profit groups and relief agencies to remove satellite photos of the affected areas from their websites (see *Nature* 437, 1072–1073; 2005). But after a brief public outcry, the authorities quickly relented. "The government and whole country have been awakened," says Umar Farooq, director of the Institute of Geology at the University of the Punjab in Lahore, who has criticized the government for its secrecy.

Chaudhry adds that the government has also begun establishing cooperative agreements with the United States, China, Iran and Japan. These would allow for the exchange of data and equipment, among other things. But Pakistan has yet to establish any working relation with India, which shares many of the same geological faults (see page 1).

Such openness is a positive step, says Julian Bommer, a seismologist at Imperial College in London. But to build an effective model of seismic risk, he says, "you need to supplement microseismic information with the earthquake history". Knowing the size and frequency of quakes along individual faults gives seismologists a sense of what those faults can produce. Such information, in turn, is a major ingredient of the complex formulae that determine earthquake risks and building codes.

Historical records of earthquakes in the region are spotty at best, Bommer says. To compensate, geologists must travel to individual faults and conduct trenching surveys, which analyse soil for signs of past earthquakes. Trenching is a rough way for geologists



The Kashmir earthquake revealed the inadequacy of Pakistan's building codes and seismological data.

W. THOMPSON

Portrait of the seismologist as a young woman

Finding a seismologist in Pakistan is nearly as difficult as finding a woman working outside the home. So the fact that Mona Lisa (a print hangs in her apartment) is both makes her an anomaly and a symbol of change.

The 33-year-old received her PhD in earthquake seismology in January from Quaid-i-Azam University in Islamabad, the first full PhD in seismology that a Pakistani university has ever produced, according to colleagues.

She was born and raised in Islamabad. Her father was a tax officer with the finance ministry and her mother a schoolteacher with a keen interest in the natural world. "My mother had a strong aptitude towards the sciences," Lisa says. "Unfortunately, she didn't pursue graduate-level

studies, so she pushed us instead."

By 'us', Lisa means herself and her 11 siblings. Collectively, the family could now be a university department all on its own: as well as Lisa, five brothers hold PhDs in physics, one sibling has a master's, and two more are pursuing PhDs.

Lisa too originally planned on being a physicist, but when she graduated from Government College in Islamabad, she found herself gravitating towards geology instead. "I liked the variety of courses," she says.

Being a woman in Pakistan's male-dominated society made things difficult. Often the male students would mock Lisa and the other female student in the programme. One time, when the class visited a fault outside of



Islamabad, the men refused to stand on the same side of the fault as the women. "Everybody was discouraging us, but by the second semester I stood first in the class," she recalls with a smile.

Lisa has now completed her thesis, a detailed seismic chart of

faults in the region of Kashmir where last October's earthquake struck. The work, which involves measurements and computer modelling, impressed reviewers such as Frank Roth, a seismologist at the National Research Centre for Geosciences in Potsdam, Germany. "She's really hard-working," he says. The only weakness in her thesis, he says, is a lack of detailed data about the region's seismic history. "But I wouldn't say that's her fault," he adds.

Lisa has now secured a professorship at Quaid-i-Azam University, and she hopes to use her position to start a new seismology programme. "After the earthquake, many students are interested in postgraduate work," she says. "I have strong hopes for training them."

G. BRUMFIEL

to get a sense of the timing and size of earthquakes along a particular fault.

Outsiders alone cannot survey Pakistan's hundreds of faults, says Bommer, who has conducted earthquake-hazard assessments in Greece, Turkey and Iran. "My experience," he says, "is that stuff that's done by flying in and out is often insufficient." In short, a comprehensive assessment of Pakistan will require homegrown talent.

Unfortunately, that talent pool is currently small. The nation of 160 million is home to perhaps half a dozen researchers with PhDs in seismology and, to date, Pakistan's many universities have produced just a single seismology PhD on their own, awarded in January (see 'Portrait of the seismologist as a young woman'). The meteorological department has only 15 employees with training in the field, and none have PhDs, says Chaudhry. One employee is now working on a PhD in China.

The problem is that geology, like so many things in Pakistan, has been driven by the country's economic needs, says Ali of Quaid-i-Azam University. "Financially, people did not see a future in the field of seismology," he says. "There were no teachers available, and most of the geophysicists preferred working for an oil company."

This, too, may be changing. At the University of Peshawar, the National Center for Excellence in Geology is ramping up its expertise in seismic studies, says Asif Khan, the centre's director. After the October quake, more students are showing an interest in seismology, and the University of Peshawar recently opened an earthquake-engineering programme. With similar efforts afoot in Lahore and Karachi, Khan predicts that "within three to four years we will have a good number of researchers".

That effort is being complemented by a mammoth investment in higher education. According to Atta-ur-Rahman, chairman of Pakistan's higher education commission, the higher-education budget has risen 1,500% in three years to roughly US\$190 million. Much of the money will go towards better university salaries to lure well-trained Pakistani scientists from foreign and industrial jobs.

On top of that, in June 2005 the meteorological department received a government grant, of US\$3.5 million for the first phase of a new national seismic network, a response to the

"Financially, people did not see a future in seismology. Most of the geophysicists preferred working for oil companies."

— Khawaja Azam Ali

2004 Indian Ocean tsunami. The network will include broadband seismic stations in Quetta, Karachi and Peshawar, plus regional centres in 14 other cities. In total, it will include more than 100 new monitoring stations. "With this network," says Chaudhry, "we should be able to completely analyse earthquakes in two to three minutes."

Ultimately, however, data and assessments can save lives only if coupled with a massive public-education campaign. That's no mean feat in a country where just over half the population is illiterate.

At present, says Bommer, "Pakistan doesn't even have a seismic building code, which is astounding". As is often the case with major earthquakes, in October most people were killed when shoddily constructed buildings

collapsed. A number of simple regulations would make buildings safer, says Bommer. For example, concrete should be mixed with pure sand instead of mixtures of soil and sand, and steel reinforcing rods holding up roofs should be joined to the frame of the house to prevent it from collapsing during shaking.

But there is little money to make such improvements. Resources have been stretched near to breaking point in the aftermath of the October quake — rebuilding is expected to cost \$3.5 billion. The government's Earthquake Rehabilitation and Reconstruction Authority is overwhelmed by the needs of the millions of people left homeless by the quake, says Mohyuddin Wani, an official with the authority.

Wani says that a new building code is being drawn up in Islamabad, based in large part on assessments by Turkish and Chinese seismological teams that visited Pakistan last November. But whether those codes can be properly distributed and enforced before reconstruction begins seems doubtful. "Pressure is building to begin reconstruction immediately," he says.

In spite of the challenges, many now believe that Pakistan might be on a course to cope with its seismic troubles. But Bilham worries that the fresh efforts to publicize data, educate more geologists and provide new instrumentation could falter as the October disaster fades from the public memory. He argues that the country needs a long-term commitment to seismic safety, as local governments in the western United States have done. "People in Pakistan are just as clever as people in California," he says. "They just haven't yet had the opportunity to deal with the problem."

Geoff Brumfiel is Nature's physical science correspondent.

THE ICEMAN OF SVALBARD

Some say that life began in fire. Hauke Trinks thinks it began in ice, and is bent on taking the hard route to prove it. **Quirin Schiermeier** tells the Arctic adventurer's tale.

Hauke Trinks is obsessed with ice, specifically sea ice — ordinary frozen sea water. What others see as a frozen wasteland, this 63-year-old German physicist sees as the kind of place where the first chemical steps that led to life may have occurred.

Trinks has undertaken daring Arctic expeditions to investigate whether, in the freezing conditions, the right mix of chemicals produces RNA molecules; many think these could have been the basis of early life. Recent lab experiments seem to lend support to Trinks's hypothesis that RNA molecules could have got their start in the frigid, rigid molecules of sea ice.

Trinks's approach to the problem is unusual. In 1999, the former president of the Hamburg University of Technology sailed alone for a year in a fjord in the Norwegian archipelago Svalbard, and set up a floating research laboratory on the boat.

His adventures there sound like a Jack London novel. A polar bear tore apart his rubber dinghy. When he ran out of food, he switched to dog food and hunted seals. When he got toothache, he pulled out the culprit with pliers — with plenty of rum serving as anaesthetic.

Despite the hardships, two years later, Trinks wanted to go back. The governor of the archipelago forbade him to go alone, so this time he took along former librarian Marie Tièche, whom he had met one night in a Svalbard pub.

For several months in complete darkness, Tièche held the fort while Trinks studied the physical and chemical properties of sea ice. "It's a rather tough job out there — you spend most of your time and energy just trying to survive," he says.

The sea ice is a far cry from the toasty environments, such as hydrothermal vents or warm ponds, that others have proposed as possible locations for the origin of life. But some scientists agree with Trinks: ice could be just right. Complex RNA molecules often degrade at high temperature, and liquid water is thought to be too reactive to support the chemical evolution of RNA. A frigid environment, possibly on an ice-encrusted 'snowball Earth', could therefore be more hospitable. Cold could even facilitate the chemical reactions needed for nucleotides — the building-blocks of RNA — to link together into RNA molecules. Perhaps minerals



Cool experiment: Hauke Trinks combines science with a thirst for adventure.

trapped in the sea ice, or surface effects of ice crystals, assist the reactions, says Gerald Joyce, a molecular biologist at the Scripps Research Institute in La Jolla, California.

Deep frozen

By measuring the physical features of sea ice, such as its electrical properties, Trinks hopes to better understand the kind of environment in which he thinks early chemical reactions may have taken place.

But because other biological molecules permeate the ice, testing whether sea ice provides ideal conditions in which the nucleotides of RNA link up is next to impossible out in the field. To do this, Trinks joined forces with Christof Biebricher, a biochemist at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany.

Biebricher created artificial sea ice in a freezer, turning the freezer off and on every few hours to simulate the temperature fluctuations of the Arctic. He added a group of nucleotides, and waited.

A year later and a half later, Biebricher thawed and analysed the samples. He found a rich harvest of RNA, with nucleotides linked together in chains more than an order of magnitude longer than those seen in other

origin-of-life studies — including the famous experiments done at higher temperatures in the 1960s and 1970s by chemist Leslie Orgel.

Orgel himself encouraged Trinks and Biebricher to submit their paper, and its publication in October¹ caused a ripple of excitement in the chemical-evolution community.

"This is fascinating," says Donna Blackmond, a physical chemist at Imperial College London. "Studies like this really help us think about where to go next experimentally. Maybe we should look at lower temperatures."

Trinks's paper takes a long run-up, first describing the complex properties of sea ice. Sea ice is quite different from freshwater ice, he explains. It's made of salt-free ice crystals enveloped by membrane-like layers of water, highly concentrated brine, carbon dioxide bubbles and salt crystals. The network of countless channels and compartments could provide sur-

faces on which RNA molecules could assemble and grow.

Many scientists working on the problem think that life's last common ancestor, which arose more than 3.5 billion years ago, was heat-loving. But this does not disprove an icy beginning, says Alexander Vlassov, a biochemist with the biotech firm SomaGenics, based in Santa Cruz, California, and author of a recent review on the subject². The earliest RNA molecules could have started out in cold environments, before moving to warmer ones.

Still, every hypothesis about the origin of life is impossible to prove conclusively. And even if Trinks's idea stands up, it would solve only a part of the problem, says Joyce. The question of where the initial building-blocks came from remains unanswered, for example.

In the meantime, Biebricher is busy replicating and improving his experiment. And Trinks left Germany on 11 February for Spitsbergen, where he is once again taking measurements of his beloved sea ice.

Quirin Schiermeier is Nature's German correspondent.

"Studies like this really help us think about where to go next."
— Donna Blackmond

1. Trinks, H., Schröder, W. & Biebricher, C. K. *Orig. Life Evol. Bios.* **35**, 429–445 (2005).

2. Vlassov, A. V., Kazakov, S. A., Johnston, B. H. & Landweber, L. F. *J. Mol. Evol.* **61**, 264–273 (2005).

BUSINESS

WESTINGHOUSE



Toshiba goes nuclear

A Japanese giant is betting big on nuclear power — but not everyone thinks the gamble will pay off, as Kurt Kleiner reports.

When Toshiba agreed last month to pay US\$5.4 billion for Westinghouse Electric, the Japanese manufacturer placed a hefty bet that nuclear power is on the verge of a global comeback.

Toshiba paid British Nuclear Fuels (BNFL) more than five times as much as BNFL paid when it bought the Pittsburgh-based company in 1999. The high price surprised the markets, and Toshiba's shares dipped by about 3% on news of the acquisition. Some analysts questioned whether the Japanese electronics giant needs the hassle of controlling the world's largest supplier of nuclear power stations.

Westinghouse designed about half of the world's 400-odd commercially operating nuclear reactors. Last year, the company, which also sells nuclear fuel and operates and maintains nuclear plants, had sales of \$1.8 billion. In a statement, Toshiba said it expects to triple that by 2015. The company is particularly confident that it can win contracts from China, which is planning to place orders soon and complete 30 to 35 new nuclear plants by 2020, at a total cost of about \$50 billion.

However, an even greater dividend would come from Toshiba's investment if other nations, such as India, Britain and the United States, decide to resume building nuclear power stations.

According to the US Energy Information Administration — a branch of the federal government charged with predicting energy trends — global nuclear capacity will grow from 361 gigawatts now to 422 gigawatts in

2025. But that modest projected increase hides a large potential demand for new stations, as many existing ones are due for replacement.

Westinghouse builds pressurized water reactors. Favoured in China and the United States, these reactors use a contained loop of very hot water to heat a secondary cooling circuit, which produces steam to drive turbines. Toshiba already sells a different design, the boiling water reactor, in which steam from water boiled by nuclear fission drives turbines directly.

"The Chinese market is the biggest game in town if you're a nuclear vendor," says Richard Lester, a nuclear engineer at the Massachusetts Institute of Technology (MIT). "If you're a company like Toshiba, which until now has been focused on boiling-water reactor technology, acquiring Westinghouse puts you in the game in China."

Adding to the attraction for Toshiba is the Westinghouse AP1000, a new design that was certified in January by the US Nuclear Regulatory Commission. The AP1000 is a pressurized water reactor that uses 'passive' safety systems said to be simpler and more reliable than those of older reactors, which depended on pumps to cool an overheating reactor. It was such an overheating incident, at Three Mile Island in Pennsylvania in 1979, that led

utilities in the United States to abandon nuclear power for a quarter of a century.

The AP1000 instead uses a mixture of gravity, convection and condensation that will, its designers say, come naturally into play if the reactor malfunctions. Westinghouse claims that an AP1000 could safely last for three days after an accident without human intervention.

But it is unclear whether better safety features and fading memories of Three Mile Island will open the way to new nuclear plants in the United States. Despite the federal government's pledge to take responsibility for waste disposal, and to meet any liability greater than \$10 billion in the event of an accident, utilities have been slow to return to nuclear energy, deterred by controversy and the risks of delays in licensing and construction.

Still, nine US utilities have embarked on a licensing process that would allow them to build between 15 and 26 new nuclear plants in the next ten years. The applications aren't a commitment, and it is expected to be several years before any utility places an order. But they show revived investor interest in nuclear power, says Caren Byrd, an analyst with financial company Morgan Stanley in New York.

Backers say that nuclear power plants are needed to help meet global electricity demand, which is projected to almost double in the next 20 years, while reducing emissions of greenhouse gases.

Doubts about the viability of such plants persist, however. A study in 2003 by researchers at MIT concluded that nuclear power was about 20% more expensive than gas and 50% dearer than coal-fired plants. Higher fossil-fuel prices have made nuclear power more competitive since then, says Lester, who co-authored the study, adding, however, that choosing nuclear "is

by no means a clearcut decision for American power generators."

Daniel Kammen, a physicist at the University of California, Berkeley, who specializes in energy policy, says that the long-term economic prospects for nuclear power are still hazy, partly because the industry is so reliant on government subsidies.

But in the short term, China seems intent on going ahead with its reactors. And some new US reactors are likely to be built soon, says Kammen, who thinks the deal makes sense for Toshiba. "It's an opportunity to acquire technology that's already certified in the United States," he says. "It's a very clever move." ■



Toshiba boss Atsutoshi Nishida says its nuclear business will triple by 2015.

I. INOUE/AP

Top-down standards will not serve systems biology

SIR — Marvin Cassman and his colleagues, in their Commentary “Barriers to progress in systems biology” (*Nature* **438**, 1079; 2005), discuss the development of standards in systems-biology research. We agree with the need for well-curated databases, software systems that can work together to analyse such data and integrated models that can deliver the fruits of systems-based research to laboratory biologists. But we have concerns about the proposed solution, which is presented as a ‘top-down’ approach that ignores many existing and emerging standards. It seems based on false assumptions about the research community and ignores the community it is intended to serve.

Cassman and his colleagues argue that standards are needed because much software developed in research settings is not reusable by other groups of working biologists, who are not appropriately trained. But the community has many excellent quantitative scientists and software developers — and with the advent of genomics, an increasing number of physicists, mathematicians, statisticians, computer scientists and engineers have joined the ranks of biologists.

It is not a lack of training that influences software design, but the realities of developing software in a research environment where developing a professional software system is not the primary goal. As fields mature and the methodologies used to generate the data become well known and established, it is both appropriate and valuable to have standardized, easy-to-use software. But standardized approaches are not always appropriate for developing software to support new research using novel methodologies in exciting new ways.

Our collective experience, gained through the Microarray Gene Expression Data Society and the BioConductor project, clearly demonstrates that flexible systems are needed and that most initial efforts are neither well documented nor widely used. But that is not a bad thing — as science charts a particular path, the appropriate tools, if given room to evolve, do emerge and rise to the top, becoming better documented and more robust.

Even with the relatively straightforward task of assembling and annotating genome-sequencing data, computationally elegant solutions to software interoperability (such as the common object request broker architecture, or CORBA) were ultimately abandoned in favour of FASTA-formatted sequence data and tab-delimited output from various analytical tools strung together using Perl. It wasn't elegant or pretty, but it delivered what was needed in a way that sophisticated users at various locations could

replicate and adapt to suit their needs. When combined with well-engineered databases and websites to provide access, the genome projects also delivered the fruits of their work to the broader community in a form that has been extremely useful and continues to evolve. Engineering this ahead of time, particularly when the field and the tools were evolving so rapidly, quite simply would have failed.

We believe that the centralized approach proposed by Cassman and colleagues would not fare well compared with more democratic, community-based approaches that understand and include research-driven development efforts. Creating a rigid standard before a field has matured can result in a failed and unused standard, in the best of circumstances, and, in the worst, can have the effect of stifling innovation.

John Quackenbush

Dana-Farber Cancer Institute and Harvard School of Public Health, Department of Biostatistics and Computational Biology, 44 Binney Street, Boston, Massachusetts 02115, USA

johnq@jimmy.harvard.edu

Other signatories of this letter:

Christian Stoeckert University of Pennsylvania, Philadelphia

Catherine Ball Stanford University, Stanford, California

Alvis Brazma European Bioinformatics Institute, EMBL, Cambridge

Robert Gentleman Fred Hutchinson Cancer Research Center, Seattle

Wolfgang Huber European Bioinformatics Institute, EMBL, Cambridge

Rafael Irizarry Johns Hopkins School of Public Health, Baltimore

Marc Salit US National Institute of Standards and Technology

Gavin Sherlock Stanford University, Stanford, California

Paul Spellman Lawrence Berkeley Laboratory, California

Neil Winegarden University Health Network, Toronto, Canada

Taxonomy on the fly in a European web project

SIR — Donat Agosti, in Correspondence, argues that restricted access to taxonomic information impedes many countries in their biodiversity monitoring efforts (“Biodiversity data are out of local taxonomists’ reach” *Nature* **439**, 392; 2006). This problem resulted in the establishment of the Global Biodiversity Information Facility in 2001, with free online access currently to almost 90 million records (see www.gbif.net).

The European Commission is about to sign into existence a network of excellence to create the European Distributed Institute of Taxonomy (EDIT; see www.mnhn.fr/edit), an initiative of the Consortium of European Taxonomic Facilities (www.cetaf.org).

A component within EDIT will provide a freely accessible web portal to integrate taxonomic revisions on the fly, allowing users to build a resource tailored to their needs. If data are properly structured, then identification keys, distributional checklists and other ready-to-use products can be built on demand.

There is, however, a potentially negative

impact on the career development of taxonomists. Currently, performance is measured using publication metrics that rely on the copyright restrictions to which Agosti refers. EDIT is mandated to examine those restrictions and to relieve them where possible, both through abstracting information from copyrighted publications for summary access (without copyright infringement) and by developing new metrics by which institutions can assess taxonomists’ productivity.

Readers specifically interested in making taxonomic material openly available are invited to contact the coordinator (D.R.) of this particular component of EDIT.

Simon Tillier*, Dave Roberts†

*Muséum National d'Histoire Naturelle, CP 43, 57 rue Cuvier, 75005 Paris Cedex 05, France

†The Natural History Museum, Cromwell Road, London SW7 5BD, UK

dmr@nhm.ac.uk

Taxonomy: programmes developing in the South too

SIR — According to Donat Agosti in Correspondence (*Nature* **439**, 392; 2006), biodiversity data are beyond the reach of many taxonomists in the developing world. But since 1999, the State of São Paulo Research Foundation in Brazil has supported a research programme called BIOTA/FAPESP on characterization, conservation and sustainable use of the state's biodiversity (see www.biota.org.br).

In the past six years the programme, for which I am a member of the steering committee, has produced an atlas of the remnants of native vegetation, supported 75 research projects, trained about 250 post-graduate students, catalogued approximately 10,000 species and made data from 35 biological collections freely available.

In 2001, the programme launched an open-access electronic peer-reviewed journal, *Biota Neotropica* (www.biotaneotropica.org.br), for original research on biodiversity in the neotropical region. And in 2002 the programme began a venture called BIOprospecTA (www.bioprosecta.org.br), in order to search for new compounds of economic interest.

Similar initiatives are under way in Mexico (CONABIO; www.conabio.gob.mx), Costa Rica (INBIO; www.inbio.ac.cr) and Africa (BIOTA Africa; www.biota-africa.de), showing that many developing countries are aware of their responsibilities under the Convention on Biological Diversity.

Carlos A. Joly

Department of Botany, Biology Institute, State University of Campinas, CP 6109, CEP 13083-970, Campinas/SP, Brazil
cjoly@unicamp.br

COMMENTARY

Global network could avert pandemics

Good surveillance is key to responding to a bird flu pandemic. **Jean-Paul Chretien, David L. Blazes** and their colleagues propose a new network of labs modelled on existing military facilities.

Since late 2003, an avian influenza epidemic has caused human deaths in Vietnam, Thailand, Cambodia, Indonesia, China and, this year, Turkey and Iraq. So far, most or all of these human infections were acquired directly from birds; if the virus becomes easily transmissible from person to person, we could face a devastating pandemic. To avert this, we must be able to quickly identify and contain avian influenza outbreaks, in birds and people, wherever they occur. This is challenging in developing countries, which often lack basic laboratories and epidemiologic capabilities. We are concerned that surveillance is weak in those countries that need it most.

How can we improve influenza preparedness in critical, but resource-poor settings? There is an existing model, based on a network of US military overseas laboratories, that we believe is worth copying. These laboratories work with host countries and the World Health Organization (WHO) to improve detection of and response to avian influenza and other emerging infections. US Naval Medical Research Unit-2 (NAMRU-2) in Jakarta helped to detect avian influenza in Indonesia last July, and plays an important role in monitoring the disease. NAMRU-3 in Cairo, Egypt, identified avian influenza in birds in Central Asia last summer, provided on-the-ground expertise in Turkey after human cases occurred there, and identified the first human case in Iraq.

Closing down

NAMRU-2 and NAMRU-3 are two of the few remaining overseas infectious-disease research laboratories that the US military established decades ago. The world needs such laboratories now, more than ever, as platforms for sustained epidemic detection and response — for avian influenza, and as-yet-unknown diseases. The time has come to build on their experience and create a new generation of multilateral, WHO-aligned laboratories as a front-line defence against future pandemics.

During the Second World War, the Navy established NAMRU-2 in Guam to study infectious-disease threats in the Pacific. Relocated several times for strategic and diplomatic reasons, the laboratory settled in Indonesia in 1970 after helping the government



A US Army veterinarian conducts training in Kenya on collecting avian specimens for influenza testing.

to control a plague outbreak there. Since then, NAMRU-2 has operated continuously in collaboration with the Indonesian Ministry of Health and other governmental and non-governmental organizations.

The NAMRU laboratories and the Armed Forces Research Institute of Medical Sciences, a US Army-affiliated facility in Bangkok, Thailand, were each set up following joint US and host-country concerns over a specific disease, but they have since conducted broad-ranging research. With state-of-the-art capabilities, access to areas with high disease rates, and strong cooperation between US and host-country personnel, the laboratories conducted field studies that led to vaccines for hepatitis A and Japanese encephalitis, treatments for malaria, cholera, typhoid fever and leptospirosis, and other lasting contributions. The laboratories never lost sight of important local diseases that posed little threat to US troops, and retained their welcome even when diplomatic relations worsened. As testimony to this, NAMRU-3 continued operations despite a diplomatic break between the United States and Egypt after the Six-Day War in 1967.

The US military built two more infectious-disease laboratories that remain today, in

Kenya and Peru. But during the second half of the century, more labs were closed than opened. Laboratories in Panama, Puerto Rico, Brazil, Congo, Uganda, Ethiopia, Malaysia and other countries folded for various reasons, including budget cuts and changing strategic needs. In the late 1970s, the United States even contemplated closing, or turning over to civilian contractors, all of its overseas military research laboratories. That plan was dropped after vigorous opposition from leaders in tropical medicine, in universities, government and industry, and from the host countries.

Under surveillance

In 1992, public-health experts warned that the loss of overseas laboratories left the United States, and the world, more vulnerable to emerging infections¹. In response, President Bill Clinton, in 1996, directed the Department of Defense to address emerging infections as a national-security threat. The military added a mission to its research-oriented overseas laboratories: develop surveillance networks to detect emerging infections, and contain epidemics before they spread. The US Department of Defense Global Emerging Infections Surveillance and Response System (DoD-GEIS) was

M. PARKER

established to support and coordinate these efforts, among others².

Building on the existing military laboratories, DoD-GEIS developed surveillance networks that monitor host-country populations in clinics, hospitals and communities in areas where disease epidemiology was largely unknown. From the beginning, influenza, with its ever-present pandemic threat, was the top priority. Current DoD-GEIS influenza surveillance networks include patient enrolment sites in more than 20 countries in South America, the Middle East, sub-Saharan Africa and central and southeast Asia³. In several countries, including Indonesia, the networks are the WHO's only information source on circulating strains, essential for vaccine development and pandemic preparedness. Collaboration with the US Centers for Disease Control and Prevention has strengthened these networks in Indonesia and elsewhere. The laboratories also respond to outbreaks of other diseases at the request of host countries or the WHO by sending their own field epidemiologists to the scene, or by providing definitive testing. They often identify diseases where they were not previously known to occur.

Training personnel from the host country in surveillance and outbreak response is important. Graduates of outbreak courses, often Ministry of Health staff, have gone on to apply the lessons to real epidemics. And in Peru and Egypt, the military laboratories assisted host countries in developing nationwide surveillance systems, which are now under host-country control.

Fill in the gaps

We believe that these efforts have improved epidemic preparedness in many places, especially where they complement activities of the WHO and its partners. But critical gaps remain. We cannot be optimistic that a deadly and easily transmissible disease emerging in sub-Saharan Africa, or a rural part of Asia or South America, would be detected, characterized and contained before spreading. But we do have a model for improving the situation. A network of new, state-of-the-art laboratories, built upon lessons learned by the military laboratories, could help to detect and contain

epidemics in key areas worldwide (see table).

The United States may decide in the future to establish more broad-based overseas laboratories, but we should not expect one country to build and sustain the network entirely on its own. Affiliation with a single country can complicate work where relations between sponsor and host are strained. And as we have seen, even a wealthy country with vested interest in global epidemic preparedness may withdraw support for productive overseas laboratories. Multilateral support would buffer against funding fluctuations. An instructive example is the International Centre for Diarrhoeal Disease Research in Dhaka, Bangladesh. Supported by dozens of donor countries and organizations, and governed by a multinational board, this centre has conducted important research on cholera and other diseases for 28 years.

Despite multilateral support, the laboratories must address host-country needs. The longevity of existing military laboratories is owed, in part, to bilateral development of laboratory priorities. For example, the laboratories monitor diseases important to host countries that might not pose direct threats to US troops, including yellow fever (against which US forces are vaccinated), cholera and hepatitis C. The laboratories must also balance global needs for public disease notification with local sensitivities over patient specimens and information; otherwise, host countries may be reluctant to acknowledge outbreaks.

We emphasize that the multilateral laboratories would support, not duplicate, the WHO's work. Through its Global Outbreak Alert and Response Network (GOARN), the WHO connects existing institutions and networks to identify, confirm and respond to outbreaks of international importance. But it is not part of GOARN's mission to fund the construction of new laboratories. The mission of the multilateral laboratories would be to develop regional surveillance and response capacity in resource-poor regions, but they would seek recognition from the WHO. Once again, the military laboratory experience, with

DoD-GEIS as a full GOARN partner, has demonstrated the benefits of being linked to a global professional network.

Official WHO affiliation is also important when persuading countries to report surveillance results. For example, NAMRU-3 is assisting several countries in Asia, the Middle East and Africa in establishing national

influenza surveillance systems. With technical assistance comes an obligation to make information and viral isolates available to the WHO's global monitoring efforts.

Be prepared

The global threat of avian influenza has motivated wealthy countries into action. The US Congress allocated US\$3.8 billion during this fiscal year to prepare for a pandemic; some of these funds will support overseas surveillance and response activities. At the International Pledging Conference on Avian and Human Pandemic Influenza in January, co-hosted by the Chinese government, the European Commission and the World Bank, donor countries and international health organizations pledged \$1.9 billion to fight avian influenza and prepare for a pandemic, surpassing expectations. This is an important time, during which plans for systems and laboratories will be developed. We

have an unprecedented opportunity to influence global epidemic preparedness, especially in developing countries. We urgently need more capacity to do so, and in our view there is a model waiting to be adapted.

The support of the WHO, with its unparalleled network of human and technical resources, is needed for this effort. We encourage the WHO to charter a working group to flesh out the model presented here, compare it with alternatives, and judge how feasible it is.

Our proposed network is not risk-free for the sponsoring countries; uncontrollable factors can hamper overseas activities. Sponsors may feel they risk less with discrete projects than with broad-based laboratories that cover many projects and staff and require long-term support. But the next pandemic — which may or may not involve influenza — could begin anywhere in the world. Making long-term investments in laboratory and epidemiologic capabilities could help us to detect and control it. ■

J. P. Chretien*, J. C. Gaydos and J. L. Malone are at the DoD-GEIS, Silver Spring, Maryland, USA; D. L. Blazes is at the US Naval Medical Research Center Detachment, Lima, Peru.

Acknowledgements We thank David Heymann for manuscript review and discussion. The opinions are the views of the authors, and are not to be construed as official views of the Departments of the Army or Navy, or the Department of Defense. Several of the authors are military service members. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.

1. Institute of Medicine. *Emerging infections. Microbial Threats to Health in the United States* (eds Lederberg, J. & Oaks, S. C.) (National Academies Press, Washington DC, 1992).
2. DoD-GEIS www.geis.fhp.osd.mil
3. Canas, L. C. *et al. Mil. Med.* **165**, 52–56 (2000).

OVERVIEW OF PROPOSED LABORATORIES	
Who would fund them	Multiple governments, other funding organizations
What the host country role would be	Drive priority setting Contribute some of the staff
What other sponsors' roles would be	Negotiate priorities Contribute staff for multi-year positions Foster links with other specialized laboratories
What they would do	Develop surveillance networks Respond to outbreaks Train personnel (in developing and wealthy countries) Conduct research
How they would relate to the WHO	Join WHO network as partners Seek Collaborating Center status

BOOKS & ARTS

A Silent Spring for climate change?

Helping the public understand climate change is an important part of tackling the problem.

The Winds of Change: Climate, Weather, and the Destruction of Civilizations

by Eugene Linden

Simon & Schuster: 2006. 320 pp. \$26

The Weather Makers: The History and Future Impact of Climate Change

by Tim Flannery

Allen Lane: 2006. 368 pp. \$20

David S. Reay

"Blame the trees!" shrieked a British newspaper, the *Daily Mail*. "Cutting down the rain forests may actually be a way of preventing global warming!" This and other similarly ill-considered headlines marked the publication of a recent paper in *Nature* (439, 187–191; 2006) reporting methane emission from living plants. Such emissions would constitute only a minor offset to the huge carbon sink the world's forests represent, but the message given to the public was at best confused and at worst downright misleading.

Increased public awareness and understanding of climate change is vital if we are to effectively tackle this greatest of all threats, but too often the reporting of climate science is blurred by self-interest or the need for an eye-catching headline. What's urgently needed is a *Silent Spring* for climate change: a book that will do for the fight against global warming what Rachel Carson's 1962 book did to protect the environment from chemical pollution. It will need to set out the history, science and politics of climate change in a way that is truly accessible to the public while steering clear of sensationalism and vested interests. This is a lot to ask.

The Winds of Change by journalist Eugene Linden is one of two new books that take up this challenge with verve. From the start, Linden uses in-depth research and expert opinion, rather than scaremongering and exclamation marks, to make his point. In discussing the 'gears' of the earth's climate, he expertly and succinctly describes the natural cycles that control climate and the many ways they interact. He has a great knack for metaphor, with the 'hall of mirrors' that constitutes the positive and negative feedbacks in Earth's climate system being a prime example.

The subtitle of Linden's book refers to the destruction of civilizations, and he uses the imaginative conceit of a legal case in which climate change is in the dock. The prosecution's



Body of evidence: in 1991 hikers found a man who died 5,200 years ago as temperatures dropped.

argument is that climate change has not simply been a factor in the downfall of communities throughout history, but has been the causal agent — a serial killer of civilizations rather than an accessory.

There is no denying that climate has played a role in shaping human evolution and civilization. The fall of the Akkadian empire in 2200 BC, where a period of sustained drought seems to have led to the collapse of this Mesopotamian civilization, is just one of several convincing pieces of circumstantial evidence that point the finger of guilt in the direction of climate change.

In a case so outwardly intractable as to make Jarndyce and Jarndyce from Charles Dickens' *Bleak House* seem straightforward, palaeoclimatology steps in as the forensics of climate science, providing a stream of convincing witnesses in the form of climate proxies. These include ice-core and sediment records that indicate rapid changes in global climate at the same time that certain civilizations apparently went into free fall.

We even have a corpse — that of a man found at the edge of a retreating glacier in the Alps. The man was frozen where he fell some 5,200 years ago and remained in place until found by a group of hikers in 1991. The proxy record shows that, at about the same time that the man breathed his last, so did countless other organisms around the world. Biological

activity dropped abruptly along with global temperatures.

Rather than climate change being a slow-moving beast that past civilizations could easily dodge, the palaeoclimate record indicates something very different. Big changes in climate can occur on a scale of decades rather than millennia, and so would stretch even the most resilient communities to the limit. Climate, it seems, is on a switch rather than a dial.

The case for the prosecution looks strong then. However, if climate change is a serial killer then it is also a serial creator. From the Archaean era onwards, abrupt changes in climate have been scorching one branch of the evolutionary tree only to nurture a new off-shoot somewhere else. On the charge of civilization killer, the jury is still out. Climate change was certainly at the scene of the crime on several occasions, but then so were other likely suspects such as overpopulation, disease and war.

A central question posed here is: "Are we better prepared than our ancestors were to deal with rapid changes in climate?" Linden's answer is a decisive "Maybe". That we should take note of the rapid climate shifts of the past to help avoid catastrophe in the future is clear. But history can only take us so far. We have entered a new age — the Anthropocene, as Paul Crutzen called it — in which, rather than humankind and civilization being shaped

by the weather, we are shaping the weather through our greenhouse-gas emissions.

This brings us to Tim Flannery's *The Weather Makers*. Flannery is an acclaimed scientist, writer and explorer, and his book carries a ringing endorsement from Bill Bryson, so it should be a good bet for the *Silent Spring* of climate change. Flannery is a consummate science writer and weaves together fascinating descriptions of his own exploration and experiences with interviews from other researchers. On the impacts of climate on biodiversity, I've not read anything better. From the last sighting of the golden toad in 1987 — a solitary male waiting in forlorn hope for a mate that never came — to the grotesque 'hairy

seadevil', dragged up into our world from the ocean depths and slowly dying from the heat, the writing is wonderfully evocative.

Unfortunately, there are also some niggling errors. Most of these are relatively minor, such as citing incorrect figures for Kyoto Protocol targets. But there are some clangers that Flannery or his editor should really have picked up. At one point he talks of "sea water turning acidic, its pH increasing by half to one unit". This kind of error is rather, well, basic.

Such distractions aside, this is a very good book. It covers the science and the politics of climate change in an engaging and thoughtful manner. Drawing on James Lovelock's Gaia hypothesis at regular intervals, Flannery leaves

the reader in no doubt about the magnitude of the problem we are facing and the urgency with which we need to act. The Kyoto Protocol and related political machinations are largely covered well, as are the options for mitigation through renewable energy, nuclear power and increased energy efficiency.

Is this a *Silent Spring* for climate change? It's not far off. If you want the real story on climate change — its history, science and politics — you should put down the *Daily Mail* and pick up one of these two excellent books. ■

Dave Reay is at the School of Geosciences, University of Edinburgh, Edinburgh EH9 3JN, UK. He is the author of *Climate Change Begins at Home* published by Macmillan.

FILM

Big trip to the red planet

Roving Mars

An IMAX movie produced and directed by George Butler

Walt Disney Pictures: 2006

<http://disney.go.com/disneypictures/rovingmars>

Jay Pasachoff

Robotic space exploration is humanized and dramatized when it is turned into an exciting story and presented on a screen so large, and showing such detail, as to be overwhelming. The IMAX film *Roving Mars* brings the movie skills of the Walt Disney Company to NASA's Mars exploration missions. We hear and see the stories of the Spirit and Opportunity rovers as they were built, launched and operated on the martian surface.

My bedroom television, although not small, subtends an angle of about 10°, the width of my fist held at arm's length (astronomers go around measuring angular distances with their fists and fingers all the time). An IMAX movie, seen from a typical seat in the middle of the theatre, subtends an angle of about 70°. What's more, it is 50° high, enveloping you in the picture. But IMAX is more than simply a big screen — the film in the water-cooled IMAX projectors (and in the IMAX cameras) is much larger than ordinary film, having about ten times the area, so it captures much more detail. And the detail shows: the images are crystal clear. High-quality sound, with pumping subwoofers, also helps the ambience, and *Roving Mars* has an appropriately ethereal score by Philip Glass.

The film begins by crediting NASA's Jet Propulsion Laboratory (JPL) with the unique capability of building the rovers, and explains that Steve Squyres of Cornell University was chosen to head the mission. A substantial part of the film deals with building and testing the rovers, which we see in exquisite, close-up detail. We learn how flexible each rover is, with

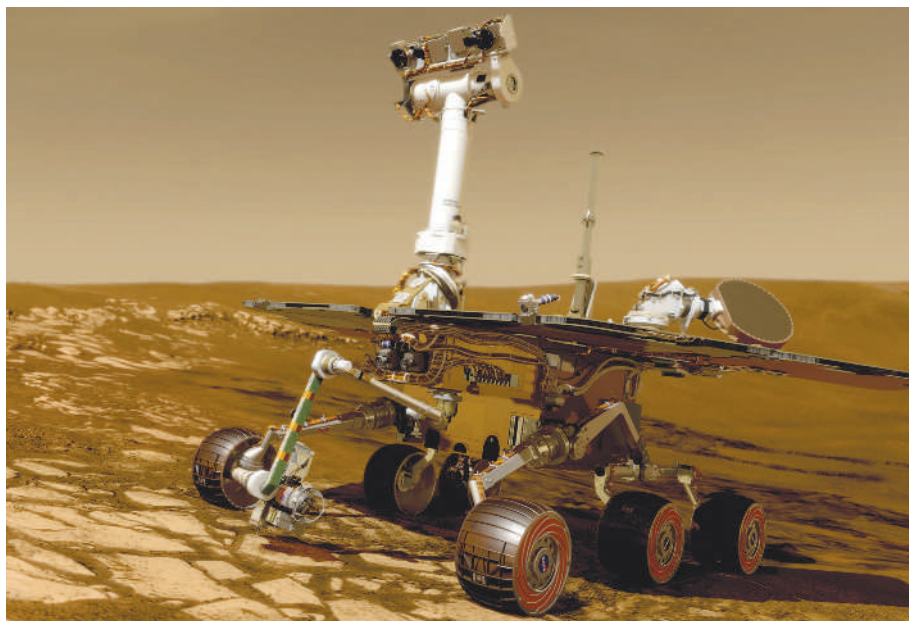
its six wheels each individually mounted and with a robotic arm that has the equivalent of shoulder, elbow and wrist motions. We see the individual devices on the arm, and how one of them, the rock abrasion tool, can cut through rock surfaces to allow unweathered surfaces to be analysed.

The movie takes us to Cape Canaveral for the launch. Dramatic high-resolution images are followed by an animation of the mission. The main problem I had with the film was distinguishing photographs from the animations, which are of exceptionally high quality; I wish the latter always had "animation" printed in a corner, or some other demarcation. And I wish the movie-makers hadn't added whooshes of sound as the later-stage rocket engines ignited, as sound doesn't travel in space.

We are shown an animation of the landing

on Mars, as no cameras hovered overhead to film the event. We see the delicate rovers cradled in spherical airbags that are cut free from parachutes and left to bounce several times, with the images animated using actual data transmitted to Earth. Real cutaways to the scientists, engineers and administrators in the JPL control room provide the spirit of space exploration. We clearly see the emotional effects of the time delay caused by the finite speed of light. And we share the elation in the control room when, after a frightening further delay, the first signals come through.

Actual film obtained on the martian surface is shown for only a few minutes at the end, but the treatment is nonetheless first-rate. The narrator discusses the evidence for past water on the martian surface, and we see types of rock that only form as sediments, such as jarosite. The movie shows, in tremendous close-up, the haematite masses known as 'blueberries' that have eroded out of the rocks — these are taken as evidence for past water on Mars. And we see some dramatic images of the planet's surface, including the crater that Spirit



The rovers' six wheels and battery of tools made them well equipped to explore the martian surface.

DISNEY ENTERPRISES



Hidden depths

The photographer Bill Curtsinger has completed 33 photographic essays for *National Geographic* magazine as part of his quest to reveal marine species in their natural settings. The photograph of the harbour porpoise above is from his book *Extreme Nature: Images From the World's Edge* (White Star, €29.20).

climbed out of, complete with tyre tracks.

Roving Mars is a model of how to draw the public into space exploration. The giant size and excellent sound make the movie especially commanding, but even the TV-resolution DVD that may soon be available would be suitable for showing in schools. A 20-page 'educator guide' can be downloaded from the official

website, which includes an excellent 90-second trailer that samples various parts of the movie.

The film can now be seen at IMAX cinemas across the United States and will be released in other countries later in 2006.

Jay M. Pasachoff is Field memorial professor of astronomy at Williams College, Williamstown, Massachusetts 01267, USA.

in communication. The contemporary importance of access, a vital performance indicator for modern museums, is highlighted.

The Renaissance *Wunderkammer*, or chamber of wonders, tended to be the private preserve of a rich or royal coterie with the means to indulge their passion for bringing together natural and artistic objects, which were often presented together in ornate cabinets. Alessandro Tosi compares the relatively unscientific *Wunderkammer* of the German tradition with the great Italian Renaissance collections of Ulisse Aldrovandi, Francesco Calzolari and others, which were almost exclusively devoted to natural history and were an important component of an emerging scientific culture.

Anna Maerker discusses the anatomical collections of La Specola in Florence and the Josephinum in Vienna, and examines the transition of the former from a cabinet of curiosities to a public museum. She also explores how the rejection by both physicians and surgeons of collections of anatomical models, which they dismissed as merely a source of entertainment for the middle classes, led to a rapprochement of these hitherto rather antagonistic disciplines. Samuel Alberti's article focuses on the concept of ownership of collections, and cites John and William Hunter as private collectors whose anatomical collections became public when they were sold to the University of Glasgow in Scotland; medical students were afforded free access while the public had to pay. Collections made and owned by learned societies formed the basis of today's great regional museums, such as those in Manchester and Sheffield, and Alberti compares the ways in which such regional museums developed.

The paper by Jonathan Simon on eighteenth-

The chamber of wonders

From Private to Public: Natural Collections and Museums

edited by Marco Beretta

Science History Publications: 2005. 272 pp. \$39.95

Kathie Way

We have all visited museums of natural history and perhaps been excited, intrigued, inspired — or even thoroughly bored — by the objects on display. But I suspect that we seldom pause to wonder how the contents of these galleries arrived in their mahogany drawer or tabletop vitrine. These essays, edited by Marco Beretta, address and offer answers to this question.

In his excellent preface, Beretta discusses the idea that the emergence of natural history as an independent discipline "was closely connected to the possession and domination of nature, rather than its contemplation". Thus it was the passion for collecting natural-history artefacts from the Renaissance to the end of the eighteenth century that drove the establishment of the discipline. The growth of exploration, the trade routes forged by the East

India Company, and the burgeoning enthusiasm for 'cabinets of curiosity', beginning in the seventeenth century, support this chronology. The 12 expert contributors to this volume provide a wealth of historiographic data drawn from late Antiquity to the nineteenth century, and examine the role of the natural-history collection through time.

Giovanni di Pasquale looks at how the Greeks and Romans documented the past through the accumulation of relics, from the ancient library at Alexandria to the established Roman museums and their collections. The suggestion that collecting relics condemns a still evident culture to the distant past is an interesting one. Susana Gómez López deals with Spanish museum collections from the fifteenth century onwards and concentrates on the link between the scientific thinking of the time and the collections in Seville, a port of entry for commodities, including slaves, from the New World. She uses a number of sixteenth-century personalities, including Spain's King Philip II, to illustrate the collecting ethos of the time and the value of collections

century mineral collections considers private collections in Paris and focuses on issues of taste and aesthetics as elements in their valuation. Beretta provides a well-researched study of mineralogical traditions from the early modern collectors who classified minerals wholly on external characters, as was common with botanical and zoological collections, to those in the latter half of the eighteenth century who used chemical characters. Much of the chapter is devoted to Antoine Lavoisier, whose work was outside the old natural-history tradition and contributed to his complex theories on the history of the Earth.

Janet Browne explores how collections may provide an entrée into a higher social group — the scientific community — and includes a particularly interesting account of how Darwin used his collections to break into academic circles and attain credibility.

Ana Carneiro looks in considerable detail at the lively discussion between the Swedish anatomist and palaeobotanist Alfred Nathorst, the aristocratic Gaston de Saporta and the Portuguese geologist Nery Delgado on what are now thought to be fossil trilobite tracks. Carneiro's thesis is that the interpretations of these objects by these three protagonists were influenced by their social and scientific backgrounds; this is interesting, but might well have been argued in rather less than 42 pages.

From a natural historian's viewpoint, these



Shelf life: by the late seventeenth century, Italy was home to some remarkable collections.

papers are far removed from the more familiar accounts of how collections, collectors and specimens contributed to our knowledge of the natural world, as they also address the largely unexplored subject of how the collections affected their collectors, an avenue of research that would reward further work. This scholarly work comprises a collection of papers that were given at a conference entitled 'From Private to Public' held in Ravenna, Italy, in June 2004. It is made accessible by the palpable enthusiasm of the authors for their subject, and is well

produced with many relevant illustrations. ■ Kathie Way is collection manager in the Division of Higher Invertebrates, the Natural History Museum, Cromwell Road, London SW7 5BD, UK, and honorary curator of the zoological collections of the Linnean Society of London.

MORE ON THE HISTORY OF MUSEUMS Cabinets For the Curious: Looking Back at Early English Museums

by Ken Arnold

Ashgate: 2006. 297 pp. £47.50

THEATRE

A defining moment for bioethics

Biblioethics: A User's Dictionary

Directed by Luca Ronconi

Teatro Vittorio in Turin, Italy

Alison Abbott

Luca Ronconi, one of Italy's most revered and provocative theatre directors, is at his most inspired when faced with the seemingly impossible. And the challenge of bringing a dictionary of bioethical terms to the stage is about as close to impossible as you can get. Yet Ronconi pulls it off in *Biblioethics: A User's Dictionary*, one of a series of five plays commissioned to promote Turin as a city of avant-garde culture on the occasion of the 2006 Winter Olympics.

Ronconi included science as one of the grand themes of contemporary humanity addressed in his series *Domani*, or *Tomorrow*. His other themes were war, history, politics and finance.

Biblioethics is not an easy play: there is no plot, just scenes that reflect on the meaning of bioethics. Ronconi disdains any technique of emotional persuasion. But he holds the attention of the audience by forcing them to participate. The theatre has been turned into a disorientating maze of small rooms separated



Biblioethics is helping the public come to terms with bioethics.

by thin partitions between which the audience has to move.

In the first room, the entire audience watches a video screen on which is projected the head of an actor reading out dictionary definitions of bioethical terms, beginning with the two that Ronconi considers central to bioethics: personal choice and informed consent. In the second room, actors playing representative characters — a clinician, a religious bioethicist,

a secular bioethicist, and male and female students — present their points of view. They speak emotionally, but their manners are deliberately detached.

The audience then divides itself between the subsequent rooms, which are dedicated to 12 different terms, including organ donation, the Nuremberg code, euthanasia and cloning. The playlets in each room run concurrently, and viewers must focus on their chosen theme, trying not to be distracted by the voices drifting across from other rooms. They realize later that they have had the chance to see

only half of the performances. Their personal choices excluded them from other options, from the sharing of other points of view.

Played out in Ronconi's hallmark monochromatic settings, the intellectual charge is animating, but the emotional charge is absent.

Biblioethics: A User's Dictionary can be seen at the Teatro Vittorio in Turin until 12 March. ■ Alison Abbott is *Nature's* senior European correspondent.

M. NORBERT

NEWS & VIEWS

EARTH SCIENCE

Megathrust investigations

Charles J. Ammon

We know the basic events of 26 December 2004: a giant earthquake beneath the Indian Ocean generated a devastating tsunami. But geoscientists are still learning about processes initiated during the earthquake.

Wherever tectonic plates meet, potential trouble exists for those who live nearby. The largest earthquakes threaten people who inhabit regions where one plate plunges beneath another. Such subduction boundaries have been gaining increasing attention as a result of new observations. Episodic tremor and slip have now been observed in several places, but are seen most clearly along the northern Cascadia subduction boundary^{1,2} off western North America. And there was the event that literally shook the world — the huge and deadly Sumatra–Andaman megathrust earthquake of 26 December 2004, which generated the large tsunami that left more than 250,000 people dead or missing³.

While geoscientists were monitoring several

subduction zones thought more likely to produce a megathrust earthquake, the 26 December event occurred in a region of relatively sparse modern instrumentation, in a place where few, if any, would have forecast such a large earthquake. Fortuitously, however, several research groups had recently completed Global Positioning System (GPS) surveys in northern Sumatra and along the Andaman and Nicobar islands. The results of these and subsequent surveys are documented by Subarya *et al.* on page 46 of this issue⁴. They present a new view of the slip associated with the Sumatra–Andaman earthquake sequence (Fig. 1) that is similar to earlier seismic and GPS results^{5,6}. But the new work is based on near-field GPS measurements and observations

of geodetic uplift that are sensitive to both the rapid fault-slip of earthquakes and the slower slip that often follows (afterslip). The sparse continuous GPS observations allow our first, albeit approximate, separation of the rapid seismic and slower subsequent slip in a giant megathrust earthquake.

Subarya *et al.* integrate observations from several sources — a near-field continuous GPS station on northern Sumatra, a set of more distant regional continuous GPS stations, several reoccupied GPS sites, and measured patterns of coral uplift. They show that most of the recent plate motion occurred along shallow parts of the plate boundary (the upper 100–150 km). The authors construct models that include a large area with more than 5 metres of slip,

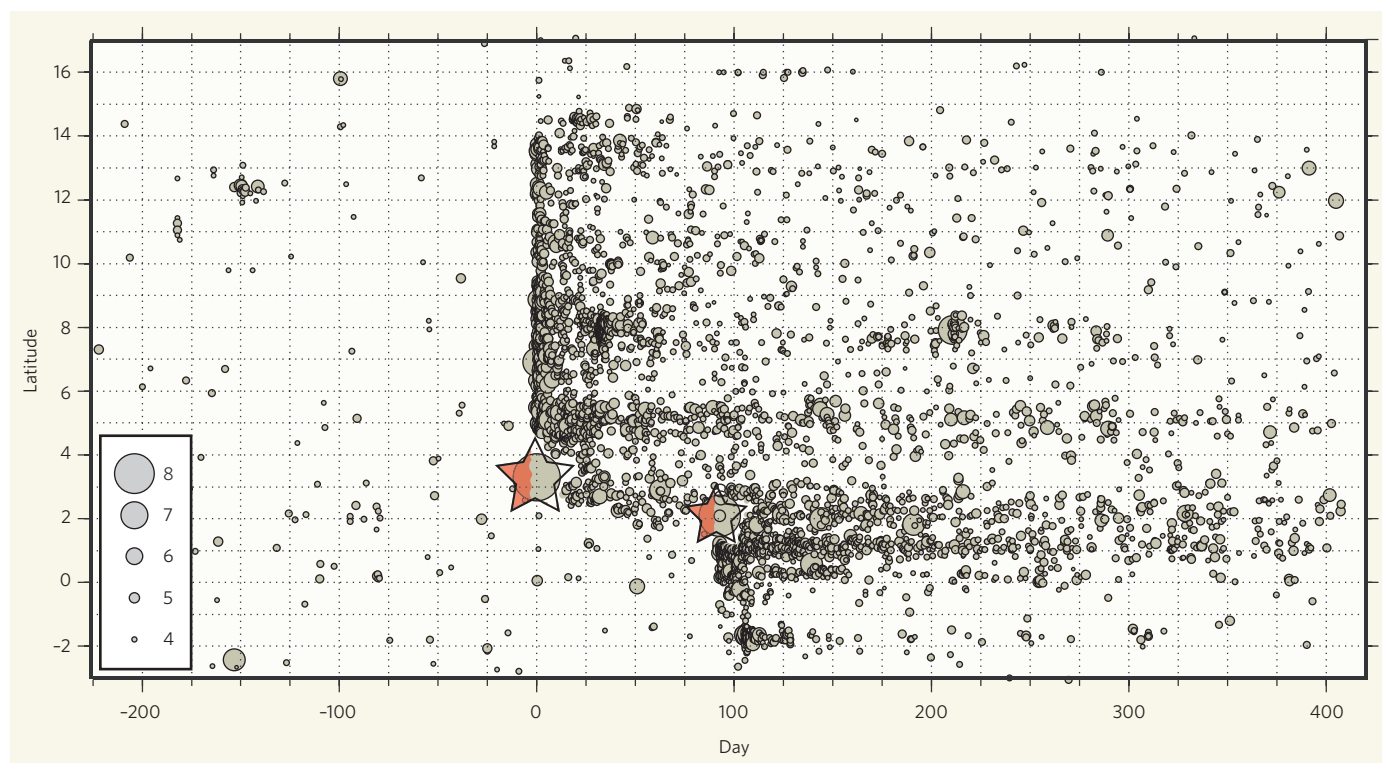


Figure 1 | Seismic activity before, during and after the great Sumatra–Andaman earthquake (including events beneath the Andaman Sea). Each circle in this space–time diagram represents an earthquake, with size increasing with magnitude as shown by the scale. The two largest events — on 26 December 2004 (day 0) and 28 March 2005 (day 90) — are identified by stars. The rupture of 26 December 2004 extended northwards for at least 1,300 km, that of 28 March 2005 southwards for 300–400 km. More than 5,000 aftershocks have been located, and activity continues across the region. (Data from ref. 13.)

peaking at more than 20–30 metres of slip off northern Sumatra (see Fig. 1 on page 47), the source region for the devastating Aceh tsunami. In the models, slip extends into the northernmost of the Andaman Islands, as surface deformation was observed as far north as Preparis Island, Myanmar. The analysis also shows that the giant earthquake (magnitude 9.1–9.2) was followed by weeks of slower deformation that accompanied thousands of aftershocks (Fig. 1). The results show that we have much to learn about how deformation is accommodated across these zones of dramatic plate convergence.

One of the most interesting aspects of the new geodetic models is the large amount of post-seismic slip, estimated at approximately 30% of that in the initial mainshock and corresponding to slip magnitudes comparable to those produced by an earthquake of magnitude 8.7. The substantial afterslip is similar to that seen with the 1960 magnitude-9.5 Chile earthquake sequence^{7,8}. But unlike the afterslip in 1960, which is thought to have occurred at greater depth than the seismic rupture, the 2004–05 afterslip seems to have been located in the shallower regions of the plate boundary, adjacent to the seismic slip that occurred abruptly on the morning of 26 December.

The emerging picture is that of a frictionally heterogeneous plate boundary that contains large regions prone to slip during earthquakes and regions that slip more slowly, although both participate in megathrust earthquake sequences. Earthquake scientists have known for some time that mapping the regions prone to sudden slip and understanding how they evolve is essential for estimating the hazards posed by a particular fault. Understanding the correlations between the regions of large seismic slip and variations in upper-plate morphology and strength^{9–11} may help to illuminate the frictional condition of the plate interface at depth. As work to integrate the available data continues, and new data on the Sumatra–Andaman plate-boundary structure are collected, perhaps we can test some of these ideas.

Many issues require deeper investigation. On 28 March 2005, a large rupture (Fig. 1) continued the sequence begun in December southeast along the Sunda arc. But why did this rupture occur 90 days after the mainshock? Subarya *et al.*⁴ identify a particular fault in the Indian Ocean floor that possibly inhibited continuation of the initial rupture in that direction. But what process changed the situation during those 90 days — afterslip-induced fault loading, migrating fluids, an expanding deformation front, or something more complicated? And what caused the rupture to terminate in the north? Did it simply run out of strain energy as a consequence of plate-motion geometry? Or was something more complicated involved — perhaps related to the subduction of a large load of sediments from the Bay of Bengal, or an undiscovered feature of plate interactions in the region? What

tectonic processes are operating in the region between this subduction zone and the Himalayan plate boundary that looms so close to billions of people?

Finally, will the sequence continue farther south? The occurrence of large earthquakes in the eighteenth and nineteenth centuries shows that large ruptures can hit this region, which includes southern Sumatra and Java (and so coastal populations of millions of people). If activity does migrate southwards, we will be better prepared with more adequate continuous GPS and seismic measurements, a result of international efforts. And at least one network is openly sharing data as they are recorded¹², which is necessary for timely and innovative integrations of the data. Perhaps these data will help in understanding the nature and implications of episodic tremor and slip before the next large subduction-zone earthquake occurs. ■

Charles J. Ammon is in the Department of Geosciences, Penn State University, 440 Deike Building, University Park, Pennsylvania 16802, USA.
e-mail: cammon@geosc.psu.edu

1. Rogers, G. & Dragert, H. *Science* **300**, 1942–1943 (2003).
2. Dragert, H., Wang, K. & James, T. S. *Science* **292**, 1525–1528 (2001).
3. www.usaid.gov/locations/asia_near_east/tsunami/trm.html
4. Subarya, C. *et al.* *Nature* **440**, 46–51 (2006).
5. Ammon, C. J. *et al.* *Science* **308**, 1133–1139 (2005).
6. Vigny, C. *et al.* *Nature* **436**, 201–206 (2005).
7. Linde, A. T. & Silver, P. *Geophys. Res. Lett.* **16**, 1305–1308 (1989).
8. Barrientos, S. E. & Ward, S. N. *Geophys. J. Int.* **103**, 589–598 (1990).
9. Wells, R. E., Blakely, R. J., Sugiyama, Y., Scholl, D. W. & Dinterman, P. A. *J. Geophys. Res.* **108**, 2507 (2003).
10. Song, T.-R. A. & Simons, M. *Science* **301**, 630–633 (2003).
11. Fuller, C. W., Willett, S. D. & Brandon, M. T. *Geology* **34**, 65–68 (2006).
12. www.tectonics.caltech.edu/sumatra/sugar.html
13. <http://earthquake.usgs.gov/regional/neic>

CELL BIOLOGY

Ageing nucleus gets out of shape

Hannes Lans and Jan H. J. Hoeijmakers

In certain premature-ageing syndromes, the architecture of the cell nucleus is abnormal. An animal model shows similar malformations during normal ageing, corroborating the idea that genome instability underlies ageing.

Hutchinson–Gilford progeria syndrome (HGPS) produces signs of dramatically accelerated ageing, such as early cessation of growth, baldness at the age of two, progressive degeneration of the skin, muscle and bone, and often fatal atherosclerosis (arterial plaque build-up) in childhood. Like many other premature-ageing syndromes, HGPS does not recapitulate all aspects of ageing — for example, patients show no neurodegeneration or cancer predisposition. HGPS and several other degenerative disorders are linked to defects in proteins that maintain the shape and organization of the nucleus, hinting that deteriorations in nuclear architecture may underlie some features of ageing. This notion is borne out by Haithcock *et al.*¹, who report in *Proceedings of the National Academy of Sciences* that progressive malformation of nuclei occurs with age in a classic animal model of ageing.

HGPS is associated with mutations in the nuclear protein lamin A (refs 2, 3). Lamin proteins polymerize to form an intranuclear scaffold known as the lamina, particularly around the edge of the nucleus. This lamina supports the nuclear architecture and helps to organize nuclear processes such as DNA and RNA synthesis^{4,5}. A remarkable diversity of degenerative human disorders is linked to different mutations in lamin A: so far, 11 so-called laminopathies are known, several of which involve premature ageing (Table 1). The common HGPS mutation results in a deletion of

50 amino acids from the protein, and gradual accumulation of the truncated, abnormal lamin A probably affects lamina function. The cells of HGPS patients show progressive abnormalities in nuclear shape, including a folded nuclear envelope and loss of peripheral heterochromatin — the densely packed DNA that is normally found in a dim rim underlying the nuclear lamina^{3,6}. This link between mutated lamin A and premature ageing does not, however, prove a causal role for nuclear architecture in normal ageing. In fact, the relevance of premature-ageing syndromes to normal ageing is highly controversial⁷.

To find out whether changes in nuclear shape are associated with normal ageing, Haithcock *et al.*¹ turned to an animal model often used in ageing research: *Caenorhabditis elegans*, a nematode worm that is well known for its extreme-longevity mutants. The mean lifespan of this nematode is only two to three weeks, but it can be extended more than five times by downregulating the insulin/IGF-1 signalling pathway. This pathway controls metabolism and influences ageing rate in all organisms investigated, including mammals⁸. To visualize the nuclear lamina, the authors engineered a worm in which the single gene for lamin was fused to a gene encoding green fluorescent protein. The fluorescent lamin protein revealed age-related, progressive changes in nuclear lamina morphology in all cells except neurons and gametes. Whereas in young

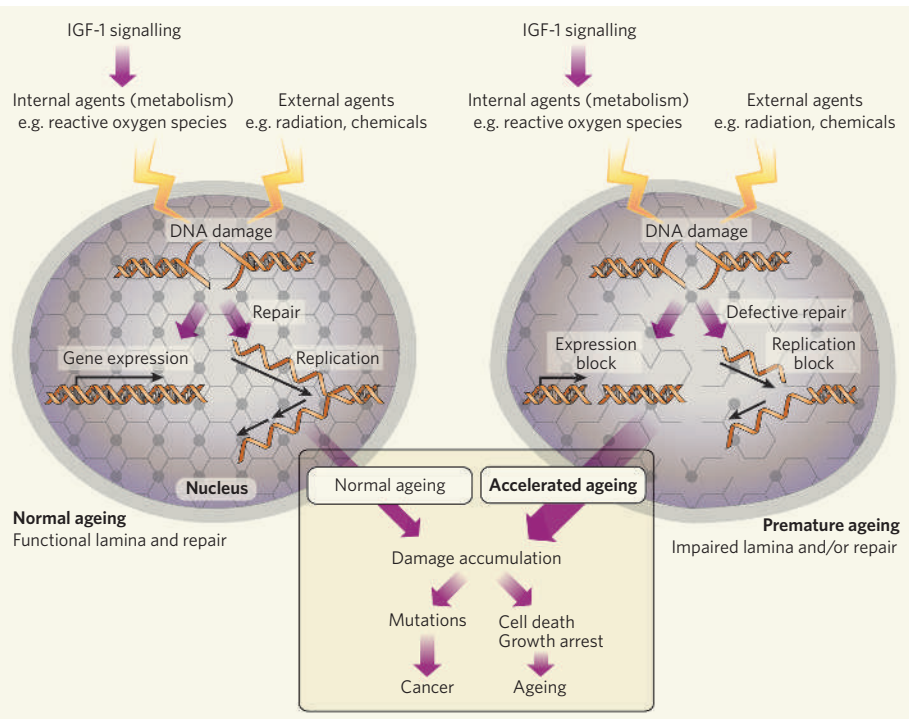


Figure 1 | Model linking normal and premature ageing with nuclear integrity and genome stability. Internal and external agents continuously damage DNA, creating a variety of harmful lesions. During normal ageing, most damage is efficiently repaired by various DNA-repair pathways, several of which are connected with gene expression and replication. Occasionally, damage leads to mutation, cell death or growth arrest, eventually causing cancer or ageing (depending on the type of damage and the pathways or cell types affected). In most premature-ageing syndromes (Table 1), DNA lesions cannot be efficiently repaired because of dysfunctional pathways for repair or damage signalling, or because the nuclear lamina is abnormal. The nuclear lamina indirectly affects genome stability because transcription and replication, and consequently also repair, take place in association with the lamina. Inefficient repair allows rapid accumulation of DNA damage, accelerating the onset of cancer or ageing.

animals the lamin protein was properly distributed around the nuclear periphery, with ageing its distribution became progressively irregular. Concomitantly, the nuclear shape changed and peripheral heterochromatin disappeared. Variations between animals and among cells underscored the stochastic nature of the changes.

Notably, the kinetics of the nuclear alterations seemed to be under the control of the insulin/IGF-1-like signalling system: the onset of severe nuclear morphological change was delayed in two long-lived *C. elegans* mutants, and it was correspondingly advanced in two short-lived animals. So, the nuclear changes seem to be connected with the normal ageing process. Although still correlative, these intriguing findings are highly reminiscent of progressive nuclear alterations found in cells of rapidly ageing HGPS patients⁶ and in cells of naturally aged mammals⁹. This supports the parallel between ageing — whether normal, accelerated or delayed — and malformation of nuclei.

How can the accumulation of abnormal lamin A contribute to ageing? There are several hypotheses^{4,5}. Structurally, lamin A

dysfunction might make the nucleus more vulnerable to mechanical stress. Although appealing, this cannot be the only explanation because not all tissues affected in HGPS patients are subject to abnormal levels of mechanical stress (for example, fat tissue). Alternatively, lamin A interacts with a diverse array of gene regulatory factors, including pRB, which controls entry into the cell-division cycle and the balance between cell division and specialization. Accumulation of truncated lamin A may mean that its binding partners end up in the wrong place and disrupt their regulatory function, thereby promoting permanent growth arrest (senescence).

We would like to propose another option. Lamin A indirectly influences genome stability because the nuclear lamina is involved in processes such as replication and gene expression¹⁰, and these processes are intimately linked with the signalling that follows DNA damage and DNA repair. Consequently, replication-associated repair (such as homologous recombination and mismatch repair), damage-tolerance mechanisms (such as translesion synthesis) and DNA-damage signalling may be compromised when the nuclear lamina is dysfunctional. Similarly, the transcription of genes is linked with the lamina and transcription-coupled repair is associated with the nuclear matrix¹¹. This repair pathway promotes cell survival by allowing recovery of gene expression blocked by damage in the DNA template.

DNA damage inevitably occurs with time; for example, because oxidative respiration continually generates highly reactive oxygen species. So, the following sequence of events may be envisaged (Fig. 1). Disruption of the nuclear lamina compromises several DNA

repair pathways. This triggers cell death and senescence upon damage induction, thereby promoting ageing. In support of this idea, there are signs of genome instability in HGPS cells and mouse mutants¹². Additionally, mice with abnormal lamin A exhibit spontaneous activation of p53 protein, the guardian of genome integrity¹³. This would add the progeroid laminopathies to the growing list of other progeroid syndromes — which lead to accelerated ageing — that are all linked with impaired genome stability (Table 1).

This scenario may explain the connection between lifespan extension and insulin/IGF-1 signalling and caloric restriction, because oxidative metabolism is a substantial cause of DNA damage: the lower or more efficient oxidative respiration is, the fewer harmful reactive oxygen species are produced, and the longer the lifespan is. This could account for

Table 1 Progeroid syndromes associated with impaired genome stability*

Syndrome (affected genes)	Affected process
Cockayne syndrome (CSA, CSB)	Transcription-coupled DNA repair
Cerebro-oculo-facio-skeletal syndrome (CSB, XPG, XPD)	Transcription-coupled and global genome nucleotide excision repair
Trichothiodystrophy (XPB, XPD, TTDA)	Transcription-coupled and global genome nucleotide excision repair
Xeroderma pigmentosum + Cockayne syndrome (XPB, XPF, XPD, XPG)	Transcription-coupled and global genome nucleotide excision repair
Xeroderma pigmentosum + DeSanctis-Cacchione syndrome (XPA, XPC, XPD)	Global genome nucleotide excision and transcription-coupled repair
Ataxia telangiectasia (ATM)	DNA damage response
Nijmegen breakage syndrome (NBST)	DNA damage response and repair
Bloom syndrome (BLM)	DNA repair and recombination
Werner syndrome (WRN)	DNA repair and recombination
Fanconi anaemia (FANC genes, BRCA2)	DNA crosslink repair
Dyskeratosis congenita (DKC1, TERC1)	Telomere maintenance
Hutchinson-Gilford progeria syndrome (LMNA)	Lamina function
Atypical Werner syndrome (LMNA)	Lamina function
Restrictive dermopathy (LMNA, ZMPSTE24)	Lamina function
Seip syndrome (LMNA)	Lamina function

*Mitochondrial DNA disorders that lead to premature ageing might also be considered to be caused by genetic instability.

why screens for mutants that have extended lifespans have never yielded genes associated with DNA repair. Apart from the fact that improving repair by single mutations is difficult, it would be impossible to improve many distinct repair systems simultaneously. Only numerous small evolutionary steps over long periods of time would eventually result in longer lifespan. However, the rate of metabolism is under the control of the single IGF-1 pathway, explaining why single mutants that reduce or enhance the rate of metabolism in general can retard or accelerate ageing, respectively, by regulating the cause of DNA damage.

Because genome stability *in toto*, including nuclear lamina function, covers almost all ageing features displayed in the wide spectrum of progeroid disorders, it is currently the most consistent parameter to underlie ageing. Thus, accelerated-ageing syndromes may be as significant for insight into normal ageing as cancer predisposition conditions have been for

understanding cancer. Similarly, DNA damage may be as relevant for ageing as it is now recognized to be for cancer.

Hannes Lans and Jan H. J. Hoeijmakers are at the MGC-CBG Department of Cell Biology and Genetics, Erasmus Medical Center, PO Box 1738, 3000DR, Rotterdam, The Netherlands.
e-mail: J.Hoeijmakers@erasmusmc.nl

- Haithcock, E. *et al. Proc. Natl Acad. Sci. USA* **102**, 16690–16695 (2005).
- De Sandre-Giovannoli, A. *et al. Science* **300**, 2055 (2003).
- Eriksson, M. *et al. Nature* **423**, 293–298 (2003).
- Worman, H. J. & Courvalin, J. C. *Int. Rev. Cytol.* **246**, 231–279 (2005).
- Smith, E. D. *et al. Mech. Ageing Dev.* **126**, 447–460 (2005).
- Goldman, R. D. *et al. Proc. Natl Acad. Sci. USA* **101**, 8963–8968 (2004).
- Miller, R. A. *Science* **310**, 441 (2005).
- Kenyon, C. *Cell* **120**, 449–460 (2005).
- Johnson, J. E. in *Ageing and Cell Structure* Vol. 2 (ed. Johnson, J. E.) 37–88 (Plenum, New York, 1984).
- Stein, G. S. *et al. Trends Cell Biol.* **13**, 584–592 (2003).
- Kamiuchi, S. *et al. Proc. Natl Acad. Sci. USA* **99**, 201–206 (2002).
- Liu, B. *et al. Nature Med.* **11**, 780–785 (2005).
- Varela, I. *et al. Nature* **437**, 564–568 (2005).

NANOTECHNOLOGY

How clean is too clean?

Ulrich Gösele

Silicon nanowires could form the building-blocks of future electronic devices, but under ultra-clean conditions, regulating their growth is difficult. Is the strictly controlled environment the problem?

At the heart of the success of microelectronic gadgets, from laptop computers to mobile phones and iPods, is the fact that silicon-based electronic devices keep getting smaller. In the most advanced of such devices currently in production, the smallest features that regulate the flow of current are less than 100 nanometres in size. But if the trend to miniaturization of these features is to continue — say, to scales below 10 nanometres — a new basic unit for electronics, other than the conventional silicon wafer, is required. Several candidate materials exist: carbon nanotubes, molecular switches and nanoscale silicon wires are examples. In a paper published online today, Hannon *et al.*¹ report *in situ* observations under an electron microscope of the growth of silicon nanowires in ultra-clean conditions*. Their observations might cast doubt on the suitability of such nanowires for mass production. They might, on the other hand, simply be telling us that the particular experimental conditions under which these nanowires were made were just too clean.

But first, some background. The idea of using nanowires of conventional semiconductors such as silicon or gallium arsenide as building-blocks of nanoelectronic or nanophotonic devices was introduced into mainstream research at the end of the 1990s (see,

for example, ref. 2). These silicon nanowires are generally grown by the vapour–liquid–solid (VLS) method. In this, a tiny liquid droplet of a metal, such as gold, absorbs silicon from a gaseous precursor, such as silane (SiH_4) or disilane (Si_2H_6), with such efficiency that the gold–silicon alloy droplet becomes supersaturated with silicon. This supersaturation causes a single, cylindrical silicon crystal — the nanowire — to nucleate, with the diameter of the nanowire being determined by the size of the initial gold droplet.

The droplet remains at the tip of the nanowire, which grows steadily outwards from it. If the nanowire lies on the surface of a silicon single-crystal wafer, growth occurs ‘epitaxially’, that is, with the same crystal orientation as the underlying silicon. The fundamentals of this technique for the growth of silicon and other semiconductors were understood more than 40 years ago^{3,4} for creating objects with diameters larger than 100 nanometres. But it was not until 1992 that the first electronic device on the nanoscale, based on gallium arsenide semiconductor nanowires, came to fruition⁵.

The droplets at the tips of semiconductor nanowires that grow in parallel have generally been considered to be independent. But Hannon *et al.*¹ show that, under the ultra-clean, high-vacuum conditions of their special

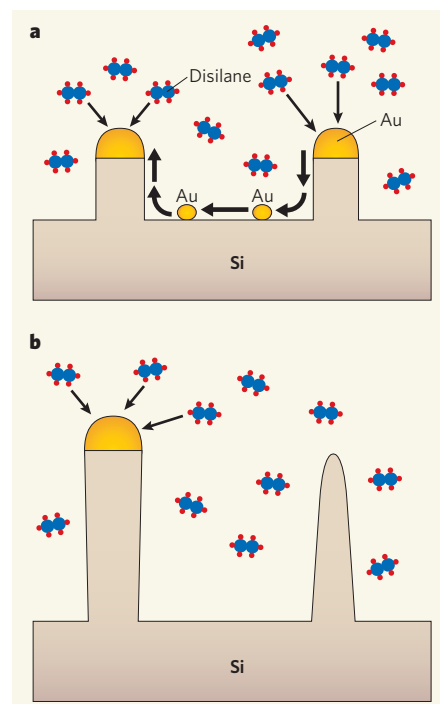


Figure 1 | Gold migration. A schematic cross-section of two silicon nanowires growing on a silicon substrate in a silicon-containing vapour (of silane, SiH_4 , and disilane, Si_2H_6) by means of the vapour–liquid–solid (VLS) method, as used by Hannon *et al.*¹. **a**, Two liquid-gold droplets are initially of slightly different size, causing a net diffusion flux of gold atoms from the smaller to the larger gold droplet through the Ostwald ripening mechanism along the silicon surface. **b**, Later, the smaller gold droplet has shrunk away completely, and its silicon nanowire has stopped growing. The implications of this effect for the mass production of silicon nanowires are potentially immense. (Dimensions not to scale; diameter changes exaggerated.)

electron microscope, larger droplets grow at the expense of smaller ones. These smaller droplets then shrink away, preventing any further nanowire growth from them. This effect, known as Ostwald ripening — sometimes jokingly referred to as the capitalistic principle — is named after Wilhelm Ostwald, 1909 chemistry Nobel laureate, who explained the effect as resulting from a decrease in total surface energy that occurs when atoms are transferred by diffusion processes from smaller to larger crystals⁶. (For a mathematical treatment of the effect, see refs 7, 8.)

Such an energy-minimizing diffusion transfer requires the efficient transport of atoms between neighbouring gold droplets. This cannot occur through gaseous diffusion above the silicon wafer because of the extremely low vapour pressure of gold; equally, the transport of gold atoms through the bulk of the silicon is also negligible. Hannon *et al.*¹ argue convincingly that the mode of transport is surface diffusion, which requires not only a high diffusivity of gold on the surface or in a thin surface layer (Fig. 1). This would fit

*This article and the paper¹ concerned were published online on 29 January 2006.

with what we know about the growth of silicon nanowires by molecular-beam epitaxy. This is a technique based on the VLS method, but in which silicon is supplied not as a gas but as a directed beam of atoms. Molecular-beam epitaxy usually requires an ultra-clean environment, and here the transport of silicon as well as that of gold occurs through diffusion on the silicon surface, not through the gas⁹.

Ostwald ripening would make the growth of ordered arrays of millions of essentially identical silicon nanowires — a prerequisite for nanoelectronic applications — exceedingly difficult. Even tiny variations in the size of the gold nucleation droplets would lead to unacceptable variations in the length and diameter of the silicon nanowires. There is, however, an escape clause mentioned by the authors that is also in agreement with our own preliminary experimental results. A tiny amount of oxygen — as is present under most technological growth conditions, but not in the ultra-clean, high-vacuum environment used in the authors' experiments — might efficiently block the diffusion path of gold on the silicon surface. This would render the gold droplets independent of each other, as has been assumed for the past 40 years.

In this way, Hannon and colleagues' results could resemble the oxygen-in-silicon story of the 1970s. At that time, manufacturers of integrated circuits were concerned that silicon crystals contained some oxygen atoms from the crystal-growth process, and pushed the manufacturers of silicon wafers to eliminate these remnants. It was subsequently observed that wafers with negligible oxygen content produced less reliable electronic devices. It turns out that oxygen in fact precipitates onto small regions of the wafers, creating traps for detrimental metallic impurities in a process known as gettering¹⁰.

Today, the oxygen content of silicon wafers is exactly specified for the best possible gettering performance. Analogously, the indirect conclusion from the fact that Hannon and colleagues¹ observe an effect under highly controlled, low-vacuum conditions that is not observed under less severely controlled conditions could be that a little added oxygen impurity — although not too much — is beneficial for silicon nanowire growth. That would be an unexpected and useful message: sometimes extremely clean is just too clean. ■

Ulrich Gösele is at the Max Planck Institute of Microstructure Physics, Weinberg 2, D-06120 Halle, Germany.
e-mail: goesele@mpi-halle.de

ECOLOGY

Corals fail a test of neutrality

John M. Pandolfi

Ecologists continue to wrestle with a central question in biodiversity studies — the prediction of species' distributions in various environments. A merger of different theories is the long-term prospect.

The most useful scientific models are those with the fewest parameters and the least-disputable assumptions. By testing real data against these 'null models', researchers are insured against over-interpreting their data. If the null model can successfully characterize the data, there is often no compelling reason to go beyond it to seek explanations.

In a new look at how species coexist in coral-reef communities, Dornelas *et al.* (page 80 of this issue)¹ have found a significant departure from the neutral theory of biodiversity² — community ecology's most recent and elegant null model. They also provide empirical support for an often overlooked component of niche theory, one of ecology's oldest and most influential explanations for the coexistence of species.

Neutral theory is one of the most exciting conceptual advances in ecology in decades. It provides a simple tool to assess the degree to which variation in the composition of natural communities can be explained solely by random demographic processes. There are only two major assumptions. The first is that every individual in a community has the same probability of birth, death, migration and speciation. Differences among species are assumed to be irrelevant when it comes to predicting community composition. The second is that the community is characterized by a zero-sum game — when one individual emigrates or dies, the space it leaves is immediately occupied by another individual. Everyone, including the theory's architect Steve Hubbell, knows that species differ in many ways. After all, not all coral species participate in the annual 'coral-reef orgy' mass spawning event; some species brood internally every month. The real question that neutral theory asks is whether or not we need to take these differences into account

to predict species-distribution patterns.

Recent tests of the neutral theory have centred on the shape of species-abundance curves³. As *prima facie* evidence against non-neutral systems, Hubbell offers the successful fitting of zero-sum multinomial functions, which meet the requirements of neutral theory, to curves from data collected in Barro Colorado Island, Panama². Since then, others have made similar comparisons using these and additional data sets, and have found both agreement with and departures from the predicted curves. More recent work has shown that most of these comparisons are weak, primarily because more than one ecological mechanism can give rise to very similar theoretical distributions, and the 'best-fit' distribution depends on what measure of 'best-fit' we use. If any model can fit, the shape that the species-abundance curves ultimately take may not be so important⁴.

In a paper that will turn our attention in a completely new direction, Dornelas *et al.*¹ describe an approach that will join an emerging next generation of tests of the neutral theory. Out go the old equivocal comparisons of the fit of different models to species-abundance distributions, and in come the new — tests that actually relate theoretical assumptions to the biological reality inherent in ecological communities^{4,5}.

Dornelas *et al.* set out to see whether local communities really do show the vast differences in species composition predicted by the demographic randomness inherent in neutral theory. Instead, they found something unexpected and initially puzzling. Coral communities along a traverse spanning part of the Indian and Pacific oceans are dramatically less similar to each other and even more variable than predicted by neutral theory. But the results are in the opposite direction to those



Figure 1 | Corals with a difference. These two shallow-water coral assemblages occur in the same habitat on the Houtman Abrolhos Islands, Western Australia, and are evidently dissimilar. Dornelas *et al.*¹ find that coral communities along their study transect are much less alike and more variable than predicted by neutral theory — which, they suggest, can be explained by environmental variance.

B. GREENSTEIN

- Hannon, J. B., Kodambaka, S., Ross, F. M. & Tromp, R. M. *Nature* **440**, 69–71 (2006).
- Morales, A. M. & Lieber, C. M. *Science* **279**, 208–211 (1998).
- Wagner, R. S. & Ellis, W. C. *Appl. Phys. Lett.* **4**, 89–90 (1964).
- Givargizov, E. I. *Highly Anisotropic Crystals* (Reidel, Dordrecht, 1987).
- Haraguchi, K. *et al. Appl. Phys. Lett.* **60**, 745–747 (1992).
- Ostwald, W. Z. *Phys. Chem.* **34**, 495–503 (1900).
- Wagner, C. Z. *Elektrochem.* **65**, 581–591 (1961).
- Lifschitz, I. M. & Slyozov, V. V. *J. Phys. Chem. Solids* **19**, 35–50 (1961).
- Schubert, L. *et al. Appl. Phys. Lett.* **84**, 4968–4970 (2004).
- Tan, T. Y. & Tice, W. *Phil. Mag.* **34**, 615–631 (1976).

predicted by many opponents of neutral theory, who generally argue for ecological building 'rules' that culminate in similar communities repeatedly being constructed in the same habitat.

The main alternative to neutral theory is niche theory, which explains relative species abundance in a community in a far more complex fashion⁶. In this theory, species have evolved adaptations to different environmental conditions and sort themselves out along 'niche axes'. These environmental conditions are both abiotic (water temperature and rainfall, for example) and biotic (competition, food supply, predation, symbiosis and so on). All else being equal, species adaptations to the environment should lead to more closely similar communities occurring over time and space than those generated by neutral theory. So, at first glance, the limited similarity in coral community composition observed by Dornelas *et al.* cannot easily be explained by traditional niche theory.

What, then, can explain this marked spatial dissimilarity in community composition? The authors suggest that environmental variability has a strong influence (Fig. 1), even though they intentionally sampled consistently similar reef habitats. Their hunch is that local communities still experience considerable fluctuations in environmental conditions. Many disturbances on coral reefs, such as cyclones, outbreaks of crown-of-thorns starfish and coral bleaching, can be quite localized, so even adjacent parts of the same reef can have very different environmental histories. Thus, the low level of community similarity might relate more to the relationship between sequential changes in communities and the time since disturbance than to random demographic processes. Although this does not show that niches have no role in species coexistence, it does place fluctuation-mediated coexistence in a stronger light⁷.

Perhaps the most instructive part of the new study is the amazingly high similarity and low variance among computer-simulated neutral-theory communities. Dornelas *et al.* tried fiddling with the parameters of neutral theory, such as migration rate and species turnover, to decrease the community similarity and increase its variance, but the simulated neutral communities remained far less variable than their real counterparts. We will have to ask some hard questions about how to interpret this result. What are the implications of a neutral model that presupposes no role for inherent differences among species, but leads to greater order and predictability in community composition than nature can provide? Will Dornelas and colleagues' conclusions hold up when the observed community similarity is much higher or less variable over different spatial⁸ or temporal^{9,10} scales? Previous authors have shown that there is less variation among communities than expected under neutral theory⁸⁻¹⁰. Dornelas *et al.* have now

demonstrated that there is too much. Neutral theory is caught in the middle. This, perhaps, shows the strength of a null model. By making clear predictions, we can explore deviations in both directions from a point of neutrality.

Dornelas and colleagues' study¹ will invigorate debate over the importance of biological details in determining the coexistence of species within communities. Their fresh and falsifiable approach leads us farther down the path to merging neutral theory and niche theory — a process that bears striking similarities to its evolutionary counterpart of successfully integrating theories concerning neutral genes and natural selection. ■

John M. Pandolfi is in the Centre for Marine

Studies and the Department of Earth Sciences, University of Queensland, Brisbane, Queensland 4072, Australia.
e-mail: j.pandolfi@uq.edu.au

1. Dornelas, M., Connolly, S. R. & Hughes, T. P. *Nature* **440**, 80–82 (2006).
2. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, 2001).
3. McGill, B. J. *Oikos* **102**, 679–685 (2003).
4. Harpole, W. S. & Tilman, D. *Ecol. Lett.* **9**, 15–23 (2006).
5. Wills, C. *et al.* *Science* **311**, 527–531 (2006).
6. Chase, J. M. & Leibold, M. A. *Ecological Niches: Linking Classical and Contemporary Approaches* (Univ. Chicago Press, 2003).
7. Chesson, P. & Huntly, N. *Trends Ecol. Evol.* **4**, 293–298 (1989).
8. Terborgh, J., Foster, R. B. & Nuñez, P. *Ecology* **77**, 561–567 (1996).
9. Clark, J. S. & McLachlan, J. S. *Nature* **423**, 635–638 (2003).
10. Pandolfi, J. M. *Paleobiology* **22**, 152–176 (1996).

DEVICE PHYSICS

Electrical solitons come of age

Thomas H. Lee

Individual packets of light energy, known as optical solitons, have long been the darlings of communications engineers. Finally, their electrical siblings are getting a look in — and could become the new favourites.

Digital communication systems seem so robust that many of us take their reliability as an axiom of modern life. That reassuring façade, however, belies a surprising fragility: in practical communication systems, impairments of a distinctly analogue nature can affect the propagation of digital pulses. Nonlinear effects, for instance, amplify the low- and high-amplitude portions of a signal differently, and so distort the signal's shape. And even without amplitude nonlinearity, the signal shape can still be distorted if the elements of different frequency of a signal wave — its so-called Fourier components — propagate at different velocities (as they do in most practical media). This kind of dispersion limits signalling to rates at which the smeared-out trailing edge of one pulse only negligibly perturbs the leading edge of the next.

The quest to mitigate effects that distort the shape of digital signals has been an obsession for communications engineers. Their efforts have paid off handsomely: modern communication channels using optical fibres can sustain data rates greater than 50 gigabits a second over a distance exceeding that between Earth's poles. That's fast enough to transmit the entire print collection of the US Library of Congress in about half an hour.

As two recently published papers^{1,2} remind us, such achievements are the result of a rare suspension of Murphy's law that anything that can go wrong will go wrong. Specifically — and remarkably — it is possible for the shape distortion produced by nonlinearity to cancel that produced by dispersion. The result is a single pulse of stable shape, dubbed a soliton.

Writing in *IEEE Transactions on Microwave Theory and Techniques*, Ricketts, Li and Ham¹ show that this fortuitous cancellation does more than simply allow the faithful propagation of digital pulses: in fact, it can be used as part of an electrical oscillator to generate pulses in the first place. And in a paper in *Physical Review Letters*, Wu, Kalinikos and Patton² describe a related system that deliberately provokes inherently nonlinear dynamics to produce chaotic soliton oscillations. Both of these systems are purely electronic; their relative ease of manufacture gives them many advantages over the 'photonic' devices, involving light waves, that currently dominate soliton research.

These modern developments ultimately trace their origins to an observation made in the early nineteenth century, when, two decades before Louis Pasteur uttered the words, the Scottish naval engineer John Scott Russell found that chance does indeed favour the prepared mind. As Scott Russell watched a barge being pulled along Edinburgh's Union Canal one August day in 1834, the rope from the barge horses suddenly snapped. He noticed that the wave produced by the prow's rapid drop onto the water's surface propagated quickly down the canal with negligible change in shape over a distance of several kilometres — as he convinced himself by pursuing the wave crest along the side of the canal on horseback.

Scott Russell was certain that the abnormally low attenuation and dispersion of this 'wave of translation' revealed principles of fundamental importance, and constructed a large water tank in his back garden to prove it.

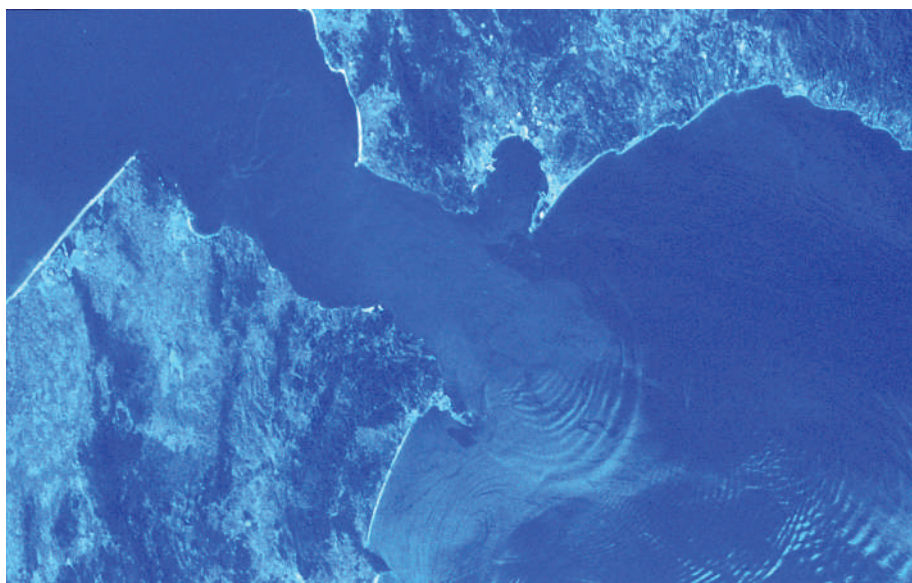


Figure 1 | Keep on waving. This image of solitons propagating out from the Straits of Gibraltar was taken by NASA's STS-41G space shuttle mission in October 1984. Lighter, fresher Atlantic sea water (which accelerates through the narrow straits from top left) compresses and upwells as it makes contact with the denser, more saline water of the Mediterranean Sea. The resulting sets of solitary waves propagate without distortion on the open sea over great distances.

Alas, he was virtually alone in this belief. Despite their appearance in various guises in nature (Fig. 1), the lack of any practical significance of such 'solitary waves' — solitons — meant that the subject would remain largely ignored for more than a century. Not until the 1960s did theoretical studies resume in earnest and reveal the conditions under which amplitude nonlinearity counteracts dispersion to permit the creation and propagation of solitons. These studies coincided with the development of lasers, whose high power density provides a practical means of teasing nonlinear behaviour out of optical fibres. It is these photonic descendants of Scott Russell's fluid solitons that have revolutionized long-distance digital communications in the past decades.

Optical soliton systems exploit the nonlinearity of an optical fibre's refractive index, under high electric fields. In contrast, electronic soliton systems use a transmission line in which the amplitude of the response to an applied pulse is itself nonlinear. Purely electronic soliton systems offer the same theoretical advantages enjoyed by their optical cousins, but have the added appeal that they are simpler to produce, as they can be manufactured using standard integrated-circuit technology.

Ricketts and colleagues¹, for example, use the voltage-dependent capacitance of conventional semiconductor junction diodes to create a discrete nonlinear transmission line. They connect this transmission line around an amplifier to make a closed feedback loop that produces an oscillating electrical signal. Ensuring the stability of an oscillating circuit requires careful control of some appropriate system parameter. In this case, the authors adjust the amplifier's gain dynamically both to guarantee that oscillations start and to avoid the onset

of chaos. The result is a stable, periodic train of self-generated solitons. The short-duration, periodic soliton train produced by the oscillator could be widely deployed in communication and instrumentation technologies.

Wu and colleagues² use a topologically similar arrangement, but bring about nonlinear feedback using a ferromagnetic film made of yttrium-iron-garnet. The complex properties of this material allow a rich variety of nonlinear behaviour, ranging from stable-amplitude oscillations to chaos. Rather than ensuring stable oscillations, Wu and colleagues operate their oscillator in a chaotic regime. The chaotic nature of the signal readily scrambles a message over a large bandwidth, and so reduces the probability that it can be detected or interfered with. Meanwhile, synchronizing a chaotic transmitter with its intended receiver permits unambiguous decoding of the original message³. Much effort is currently going into such chaotic, covert communication schemes. The convenience of having such systems in a flexible electronic form using electrical solitons will do much to accelerate their development.

From chance observation of a summer canal-side ride, to linchpin of new communications technologies: John Scott Russell would no doubt have been delighted to see his belief in the importance of solitons so vindicated. ■ Thomas H. Lee is in the Center for Integrated Systems, Stanford University, 420 Via Palou Mall, Stanford, California 94305-4070, USA. e-mail: tomlee@smirc.stanford.edu

1. Ricketts, D. S., Li, X. & Ham, D. *IEEE Trans. Microwave Theor. Tech.* **54**, 373–382 (2006).
2. Wu, M., Kalinikos, B. A. & Patton, C. E. *Phys. Rev. Lett.* **95**, 237202 (2005).
3. Roy, R. *Nature* **438**, 298–299 (2005).



50 YEARS AGO

Marine biologists have been slow to recognize that the smaller phytoplankton elements which will pass through the finest nets are of great importance in the productivity of the oceans. In order to get some quantitative data on this subject for eastern Australia (lat. 33–34° S.), we have recently conducted experiments in which 44 gallons of sea-water were pumped into a drum and filtered through a phytoplankton net having 170 meshes per inch. The filtrate was collected and the particulate matter was spun out of 2-litre aliquots using a continuous centrifuge running at 13,000 rev./min... Both the net fraction and the centrifugate were examined microscopically, using the fluorescence of chlorophyll, to count the photosynthetic organisms... It has been found that the chlorophyll content of the centrifugate is 25–3,000 times as great as that of the net plankton, while the counts of organisms are 10–10² as high again... It is proposed to continue these studies over an annual plankton cycle, in an estuary, and in the open sea. From *Nature* 3 March 1956.

100 YEARS AGO

A magnificent fireball was seen by many persons in the north of England on the evening of January 27 at 8h. 33m... Mr. H. Beckwith, at Hull, observed the meteor travelling horizontally between the "square" of Ursa Major and the Belt of Orion, while at Cheadle, Miss Blagg noted the path as just above ζ Leonis. Mr. R. Felton, at Patrington, estimated the brightness as quite equal to that of the full moon... The meteor gave a very brilliant flash near its end point, and the suddenness of its apparition startled many people... the height of the meteor was from about 59 to 45 miles over the North Sea immediately east of the Lincolnshire coast. The length of the observed path was approximately 42 miles, and probable velocity of the object 24 miles per second... In recent years fireballs have been very numerous this month, and especially at the epochs about January 9 to 13 and 24 to 29. From *Nature* 1 March 1906.

50 & 100 YEARS AGO

BEHAVIOURAL ECOLOGY

Push on the marching crickets

Stephen J. Simpson and colleagues have a gory story to tell (*Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0508915103; 2006). They investigated the factors that contribute to the mass movements of flightless Mormon crickets (*Anabrus simplex*), which inhabit western North America. Under certain conditions, these insects form into marching bands millions strong and several kilometres in length, and can walk up to 2 kilometres a day.

From their studies, Simpson *et al.* conclude that group movement is driven not only by 'pull' from in front, in the form of the search for specific nutrients (protein and salt), but by 'push' from behind — cannibalism. The crickets are themselves "walking packages of protein and salt". So unless an individual is in the forefront of the band as it rolls into new territory, its only recourse to satisfy its dietary needs may well be to make a meal of a fellow traveller. As this picture shows, Mormon crickets have a

taste for their own. It was taken near Reno, Nevada, and shows one unfortunate individual, killed by a car, being consumed by another cricket (a phenomenon that sometimes occurs on a mass scale).

To identify which nutritional components might drive mass movement, Simpson and colleagues conducted feeding-preference tests for nutrients placed in the path of a marching band. The results showed, for example, that marching crickets have a decided requirement for protein over carbohydrate, and for salt solution over water (to the extent that the insects fought for the preferred salt concentration).

Other experiments demonstrated that crickets that had had their fill of protein and salt were much less likely to be cannibalistic. And other experiments again, using crickets with compromised movement, showed that the crickets safest from cannibal attacks were those that were able to defend



themselves with their hind legs. But inactivity or reduced mobility meant that the cricket concerned was literally dead meat.

Cannibalism presumably sets a high cost on group life, so why do the bands of mormon crickets stick together? Simpson *et al.* answer this question by pointing to previous work involving the use of radiotelemetry which showed that the price of leaving the

migratory group is the high likelihood of being killed by a predator. Better, then, to opt for the protection conferred by being one of a crowd, even at the risk of being consumed by one of your own — or, as the authors put it, being a member of a cricket migratory band "is a compromise that makes the best of a seemingly very bad situation".

Tim Lincoln

D. REID/AP

CELL BIOLOGY

A break in the chain?

Keith Burridge

How chains of proteins link transmembrane cell-cell adhesion molecules to the cell's inner scaffold was standard textbook material. But recent research challenges the accepted model, opening a new chapter in the field.

Many of the adhesion molecules that enable cells to stick together are anchored to part of the cell's internal scaffold — the actin cytoskeleton. This association is usually indirect and serves to transmit tension to sites of adhesion, to cluster adhesion molecules and to provide a framework for the assembly of complexes of signalling proteins. How the adhesion molecules connect to the cytoskeleton has often been difficult to define, but one of the most widely accepted chains of attachment is that of the cadherin adhesion proteins to actin filaments through the β -catenin and α -catenin proteins, with α -catenin providing the link to actin (Fig. 1). (Indeed, 'catenin' derives from the Latin for chain.) Two *Cell* papers from the groups of W. James Nelson and William Weis^{1,2} challenge the validity of this linkage, and particularly the role of α -catenin.

Cadherins mediate cell-cell adhesion in many cell types. Much of what is known about

these proteins derives from work on E-cadherin, the type found in the epithelial cells that line ducts and cavities in the body. E-cadherin is concentrated within cell-cell contacts known as adherens junctions, and the extracellular domains of the protein hold adjacent cells together. Adherens junctions associate with a circumferential belt of actin filaments around the cell, and together these structures determine the shape of epithelial cells and maintain epithelial integrity. During tissue development, contraction of the circumferential belt can transmit tension to adherens junctions, causing cells to constrict in their apical regions, inducing the epithelium to curve or fold. This underlies the generation of structures such as the neural tube, which develops from a localized folding of the neural epithelium and eventually goes on to form the brain and spinal cord.

The first indications that cadherins might be

linked to actin filaments came from the localization of adherens junctions with bundles of actin filaments. Furthermore, cadherins partitioned with much of the cytoskeleton in crude subcellular fractionation procedures. Then, nearly 20 years ago, cadherins were shown to interact with at least two cytoplasmic proteins, α - and β -catenin. Binding studies demonstrated that β -catenin forms a bridge between cadherin and α -catenin, which in turn binds to actin *in vitro*. From these binary interactions, a chain of attachment seemed logical (Fig. 1), and this model was accepted by the field (and textbooks) with little challenge, except for the qualification that many other actin-binding proteins also bind to α -catenin³ and might contribute to this linkage.

The Nelson and Weis groups^{1,2} provide several lines of evidence that are inconsistent with this textbook linear model. They report¹ that α -catenin does not bind simultaneously to actin and to β -catenin in complex with the E-cadherin intracellular domain. They developed a system to test the binding of purified components to 'membrane patches' enriched for cadherins but stripped of the associated proteins. Although β -catenin bound to these preparations and conferred α -catenin binding, these complexes were unable subsequently to bind to either G- or F-actin.

Recognizing that additional α -catenin-binding components may be required, the

authors tested two prime candidates, vinculin and α -actinin, but these revealed little (vinculin) or no (α -actinin) binding to the cadherin–catenin complexes, and they failed to mediate actin binding to the membranes. Cell extracts, which would be expected to contain the putative missing factor, also failed to confer actin binding in this system.

Finally, the authors used FRAP, a technique that analyses the dynamics of fluorescently tagged proteins in cells or subcellular compartments, to monitor the cadherin–catenin complex and actin in adherens junctions. All the components of the cadherin–catenin complex had similar dynamics, and were mostly immobile, presumably held in the complex. By contrast, actin and several actin-binding proteins were far more mobile, arguing against a stable linkage between the cadherin–catenin complex and actin filaments.

So what is α -catenin doing instead of making a static bridge between actin and the catenin–cadherin complex? In their second paper², Nelson, Weis and their colleagues demonstrated that α -catenin exists in three forms in the cell: singly, paired with itself (a homodimer) or paired with β -catenin. The monomer preferentially binds to the β -catenin–E-cadherin complex, whereas the homodimer binds to filamentous actin (Fig. 2). Significantly, they found that the α -catenin homodimer competes with the

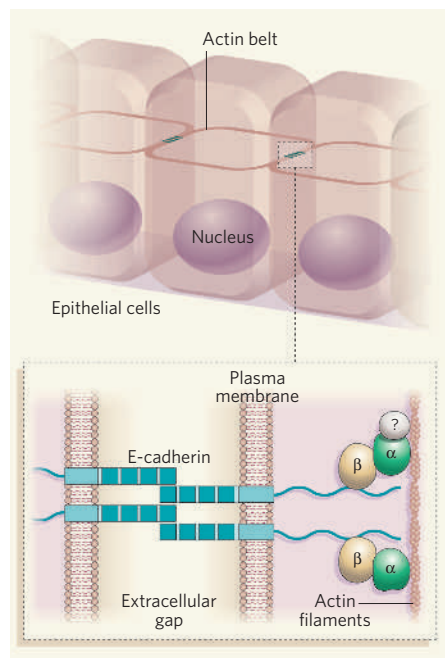


Figure 1 | The conventional model for how a cadherin links to actin filaments. The cytoplasmic domain binds to β -catenin (beige), which complexes with α -catenin (green). In turn, α -catenin provides the link to actin, either directly or indirectly through one of several other proteins (?). Work from Nelson, Weis and colleagues^{1,2} challenges this static model, instead proposing that the binding of α -catenin to β -catenin is incompatible with the simultaneous binding of α -catenin to actin (Fig. 2).

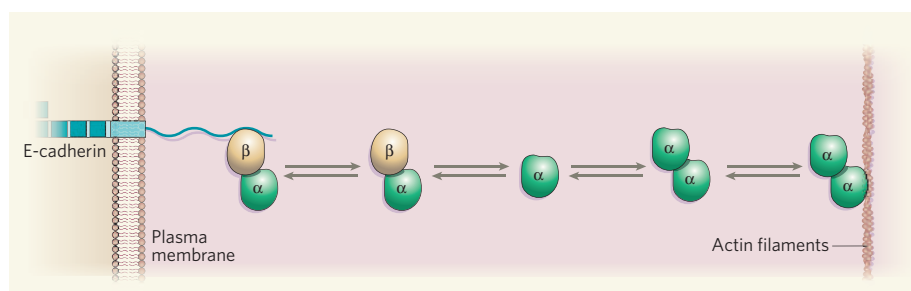


Figure 2 | The states of α -catenin. Nelson, Weis and colleagues^{1,2} find that in the cytoplasm, α -catenin can exist singly (a monomer), paired with itself (a homodimer), which binds to filamentous actin, or paired with β -catenin (a heterodimer). When complexed with β -catenin, it can associate with the cytoplasmic domain of cadherins but not with actin.

Arp2/3 protein complex for binding to actin. When bound to actin filaments, the Arp2/3 complex initiates the assembly of new actin filaments, generating a branched actin network. By preventing binding of the Arp2/3 complex to actin, the α -catenin homodimer inhibits new filament assembly *in vitro*. This inhibition requires α -catenin concentrations higher than those estimated in cells, but the authors speculate that recruitment of α -catenin to adherens junctions may provide a reservoir for the protein, increasing its local concentration next to junctions. Once there, it may contribute to the conversion of a branched actin network to bundles of parallel filaments. This must occur when adherens junctions are being established in response to cell–cell contact, although other mechanisms may also contribute to this conversion⁴.

The authors conclude that α -catenin does not link cadherins directly to actin and that the cadherin–catenin complex is not stably associated with actin. Both conclusions are provocative and demand re-evaluation of older work. The first point, for example, is difficult to reconcile with experiments by Nagafuchi, Tsukita and colleagues^{5–7} that study the function of α -catenin by fusing it or its domains directly to E-cadherin. These chimaeric constructs, particularly one that included just the carboxyl terminus of α -catenin (containing the actin-binding domain), bestowed on cells lacking E-cadherin the adhesion properties of epithelial cells. Some of these experiments implied that both E-cadherin and a chimaera containing the actin-binding domain of α -catenin were tethered to the cytoskeleton and could resist significant forces when pulled with optical tweezers⁷. These experiments, although suggestive, are not definitive, as chimaeras can exhibit unique characteristics not shared by the parent proteins, and the carboxy terminus of α -catenin can bind to other actin-binding proteins³.

The second conclusion seems to fly in the face of evidence that considerable tension can be exerted on adherens junctions. This is seen not only during development, but also in other situations; for instance, when the wound-response factor thrombin increases endothelial permeability. This occurs in part by stimulating cell contraction. The increased tension is transmitted to the adherens junctions,

changing the cell borders from a smooth to a jagged contour, with bundles of actin filaments terminating at these sites rather than running parallel to the junctions.

Can relatively dynamic linkages allow the transmission of force? With integrins (another class of adhesion molecule), tension does seem to be conveyed through low-affinity associations made between the cytoplasmic domain and intermediates such as talin. In part, this may be explained by the ‘Velcro’ principle, where many weak interactions generate an adhesion capable of transmitting large forces. Another possibility is that tension itself exposes ‘cryptic’ binding sites on α -catenin or other components, which then enhance the binding. The Nelson and Weis groups demonstrate that α -catenin does undergo conformational changes as a consequence of binding to another molecule. Thus, the interaction of β -catenin with the amino terminus of α -catenin inhibits the binding of the α -catenin carboxy terminus to actin. Could there be conditions, such as mechanical tension, that expose this actin-binding site on α -catenin while it is still bound to β -catenin?

The death knell of a research field occurs when its evolving ideas are captured by textbooks as ‘settled’ dogma, stifling further enquiry. The Nelson and Weis papers^{1,2} not only identify new functions of α -catenin but, by challenging long-accepted concepts, they should reopen a seemingly settled question and reinvigorate the field.

Keith Burridge is in the Department of Cell and Developmental Biology, and the Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA.

e-mail: Keith_Burridge@med.unc.edu

1. Yamada, S., Pokutta, S., Drees, F., Weis, W. I. & Nelson, W. J. *Cell* **123**, 889–901 (2005).
2. Drees, F., Pokutta, S., Yamada, S., Nelson, W. J. & Weis, W. I. *Cell* **123**, 903–915 (2005).
3. Kobiela, A. & Fuchs, E. *Nature Rev. Mol. Cell Biol.* **5**, 614–625 (2004).
4. Kobiela, A., Pasolli, H. A. & Fuchs, E. *Nature Cell Biol.* **6**, 21–30 (2004).
5. Nagafuchi, A., Ishihara, S. & Tsukita, S. *J. Cell Biol.* **127**, 235–245 (1994).
6. Sako, Y., Nagafuchi, A., Tsukita, S., Takeichi, M. & Kusumi, A. *J. Cell Biol.* **140**, 1227–1240 (1998).
7. Imamura, Y., Itoh, M., Maeno, Y., Tsukita, S. & Nagafuchi, A. *J. Cell Biol.* **144**, 1311–1322 (1999).

STRUCTURAL BIOLOGY

Designer labels

Stanley J. Opella

An innovative approach to the isotopic labelling of proteins places the determination of the structures of larger proteins — particularly those in solution — within reach, while improving the accuracy of the method.

With completion of the sequencing of the human and many other genomes, molecular biology is able to return to its roots to focus on, as Francis Crick¹ put it, “the description of [proteins, viruses, bacteria and chromosomes] in terms of their structure, i.e. the spatial distribution of their constituent atoms”. Significantly, this definition includes systems that range from individual globular proteins in aqueous solution or highly regular crystals, to large assemblies of macromolecules, such as virus particles and membranes, that behave like aperiodic solids².

Nuclear magnetic resonance (NMR) spectroscopy can be used on molecules in all physical states and degrees of order, and so is fundamentally capable of determining the structures of proteins in all these environments. Although powerful, the full promise of this technique has yet to be realized. In this issue, Kainosho *et al.* (page 52)³ demonstrate a method to optimize the isotopic make-up of the proteins being studied. This allows far more information to be gleaned by both simplifying and sharpening the spectra.

The potential of NMR spectroscopy for structure determination was obvious from the earliest results, when Pake⁴ showed that the distances and orientations of pairs of nuclei could be determined from the perturbations that the magnetic field of one nucleus causes in a neighbouring nucleus. The initial ¹H NMR spectrum of a protein⁵ showed that a dispersion of chemical shift frequencies results from the local spatial relationships of atoms that are not necessarily bonded to one another; thus, the chemical shift interaction provides a means to distinguish all of the atoms in a protein.

The Achilles' heel⁶ of protein NMR is the 'size problem', which has two distinct components. The first is obvious: the spectra of larger polypeptides are more complex, with more signals and consequently overlap within the same limited range of frequencies. This is compounded by the second, more daunting, 'correlation-time problem': larger polypeptides (and their complexes) tumble slowly in solution⁷, which results in broad linewidths. The correlation-time problem can be solved under certain circumstances for proteins that are immobile (on NMR timescales) by virtue of their being in the solid state or very large complexes. But the problem has so far been intractable for

larger soluble proteins in their native aqueous environment.

The past half-century of technological developments in NMR for investigating proteins has followed three parallel paths: instrumentation, experimental methods and isotopic labelling. The benefits of advances in all three areas are highly synergistic, despite their seemingly disparate nature. The selective replacement of hydrogen (¹H) with deuterium (²H) in a protein⁸ was the first real step towards the resolution of overlapping NMR spectra, preceding the advent of high-field magnets and the sophisticated magnetic pulse sequences employed today. Deuterium nuclei couple much more weakly with the external magnetic field (and with the nearby ¹H nuclei of interest), so that replacing most ¹H by ²H has a substantial effect on the properties of the remaining ¹H nuclei. In various incarnations, the incorporation of deuterium remains among the most widely used methods in protein NMR, even though, up to now, it required several difficult trade-offs.

When uniform deuteration is performed at very high levels, there are no ¹H nuclei left as sources of magnetization or the couplings that provide structural information. Intermediate levels of deuteration generate a mixture of proteins with the ²H nuclei in different positions, complicating the chemical shift patterns, and the signal intensity decreases because there are fewer ¹H nuclei in the sample.

Kainosho and colleagues³ describe a strategy for optimal isotope labelling for protein NMR that addresses these problems. Their approach goes far beyond just improving sample preparation — it is a major step forward in the attack on NMR's size problem. In addition, it provides a path to structure determination with higher spatial resolution. The authors developed protocols for 'stereo-array isotope labelling' (SAIL), which incorporates deuterium into a protein's constituent amino acids such that each carbon or nitrogen nucleus in the final protein will have at most one ¹H nucleus bonded to it, the remaining hydrogen atoms having been replaced by ²H. These amino acids are then used to make the protein in a cell-free expression system that maintains the careful positioning of the isotopes. The patterns of replacement are designed to provide the most information consistent with spectral simplification and isotopic dilution (Fig. 1 on page 52), and it constitutes a general approach to overcoming the limitations of

deuteration in a way that also enhances the contributions of ¹³C and ¹⁵N labelling for both distance and angle measurements throughout the protein. Contemporary alternatives are more limited; their focus is on specific labelling of methyl groups in a highly deuterated background^{9–11}.

The SAIL technology improves NMR spectra because it decreases the number of signals without losing sensitivity for the resonances of interest, and it produces narrower linewidths because the dilution of ¹H nuclei attenuates the line broadening. Moreover, it makes the short-range distance measurements more accurate, and, combined with easier assignment of side-chain resonances, leads to higher-resolution structures. Kainosho *et al.* demonstrate their approach by solving the structures of two proteins — calmodulin (relative molecular mass 17,000) and maltose binding protein (41,000) — at a resolution of about 1 Å. This rivals the highest resolution of structures previously determined for smaller proteins by solution NMR or X-ray crystallography.

The combination of SAIL with other NMR technologies means that protein structure determinations can be expected where accuracy is no longer limited by experimental methodology. Rather, the structures could reflect the fundamental dynamic properties of the backbone and side-chain atoms — portraying the essence of a working protein. Equivalent improvements in solid-state NMR studies can be readily envisaged.

Kainosho and colleagues' study is a significant advance in the field of protein NMR. It demonstrates that by dipping your heel in the river while sailing, the protein size problem will fade away. This will allow the study of proteins selected on the basis of their biological functions or in genuinely unbiased surveys of proteomes, resulting from its ease of automation. ■

Stanley J. Opella is in the Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0307, USA.
e-mail: sopella@ucsd.edu

- Judson, H. F. in *The Eighth Day of Creation* 88 (Cold Spring Harbor Laboratory Press, New York, 1996).
- Schrödinger, E. *What is Life?* (Cambridge Univ. Press, 1944).
- Kainosho, M. *et al.* *Nature* **440**, 52–57 (2006).
- Pake, G. E. *J. Chem. Phys.* **16**, 327–336 (1948).
- Saunders, M., Wishnia, A. & Kirkwood, J. G. *J. Am. Chem. Soc.* **79**, 3289–3290 (1957).
- Homer *The Iliad* (Penguin, London, 1999).
- Einstein, A. *Ann. Phys. (Leipz.)* **17**, 549–560 (1905).
- Markley, J. L., Putter, I. & Jardetzky, O. *Science* **161**, 1249–1251 (1968).
- Chaykovsky, M. M. *et al.* *J. Am. Chem. Soc.* **125**, 15767–15771 (2003).
- Tugarinov, V. & Kay, L. E. *J. Biomol. NMR* **28**, 165–172 (2004).
- Tang, C., Iwahara, J. & Clore, G. M. *J. Biomol. NMR* **33**, 105–121 (2005).

HYPOTHESIS

Introns and the origin of nucleus–cytosol compartmentalization

William Martin¹ & Eugene V. Koonin²

The origin of the eukaryotic nucleus marked a seminal evolutionary transition. We propose that the nuclear envelope's incipient function was to allow mRNA splicing, which is slow, to go to completion so that translation, which is fast, would occur only on mRNA with intact reading frames. The rapid, fortuitous spread of introns following the origin of mitochondria is adduced as the selective pressure that forged nucleus–cytosol compartmentalization.

The discovery of introns had a broad effect on thoughts about early evolution. Walter Gilbert coined the terms introns and exons while predicting that their utility in gene evolution might lie in exon shuffling and alternative splicing¹. Ford Doolittle suggested that the ancestral state of gene structure might be 'split' and that eukaryotes might have preserved, in the guise of exons, relics from the primordial assembly of genes via recombination in introns². James Darnell submitted similar ideas at a time when the issue in early evolution was how to generate long coding sequences from scratch³. Exonic modules provided an answer, and the introns-early view was born: the intron-bearing eukaryotic lineage was seen as an independent primordial line of descent whose genealogy went straight back to the origin of genes, as evidenced by their introns, and whose descendants included the host that acquired the mitochondrion⁴. Because introns-early proposed a modular assembly of genes⁵, it readily accommodated the discovery of catalytic RNA⁶ and the 'RNA world' concept⁷. But it also predicted exon boundaries in genes to correlate with domain boundaries in proteins⁸, and hence was abandoned when that prediction failed⁹. Yet exon shuffling and alternative splicing¹ remain as paradigms of eukaryotic gene evolution, and the effect of introns-early on theories about early evolution persists: the notions that eukaryotic introns might be direct holdovers from the origin of genes and that prokaryotic genomes were once full of introns, but became streamlined, still enjoy some currency^{10–12}.

Throughout all that, the origins of mRNA introns themselves remained an issue. Cech¹³ proposed that eukaryotic introns and their cognate spliceosomal small nuclear RNAs (snRNAs) originated from disarticulate group II introns. Cavalier-Smith¹⁴ extended that idea by suggesting that they evolved specifically from group II introns that invaded the ancestrally intron-less eukaryotic genome through the mitochondrial endosymbiont, thereby generating two predictions. First, group II introns should be found among free-living α -proteobacteria, the ancestors of mitochondria. Second, eukaryotic lineages that were then suspected to primitively lack mitochondria, such as *Giardia intestinalis* (syn. *G. lamblia*), should lack introns.

The first prediction was borne out directly¹⁵, supporting the idea that introns could originate from mitochondria^{14,16}. The mobility of group II introns in contemporary eubacteria¹⁷ and their prevalence in α -proteobacteria¹⁸ are still fully consistent with that view. The second prediction turned out to be wrong, but for an unexpected reason: *Giardia* has introns after all¹⁹, but it also has mitochondria²⁰. In fact, all of the eukaryotes that had been predicted to lack introns and mitochondria¹⁴ have both^{19–24}. So the suggestion that introns and the

spliceosome arose in the wake of mitochondrial origin could still be right¹⁴.

Here we revisit the possible evolutionary significance of introns in light of mitochondrial ubiquity. We propose that the spread of group II introns and their mutational decay into spliceosomal introns created a strong selective pressure to exclude ribosomes from the vicinity of the chromosomes—thus breaking the prokaryotic paradigm of co-transcriptional translation and forcing nucleus–cytosol compartmentalization, which allowed translation to occur on properly matured mRNAs only.

New consensus, new observations

A current consensus on introns would be that prokaryotes do indeed have group II introns but that they never had spliceosomes¹⁸; hence, streamlining in the original sense (that is, loss of spliceosomal introns) never occurred in prokaryotes, although it did occur in some eukaryotes such as yeast or microsporidia^{25,26}. An expansion of that consensus would be that spliceosomes and spliceosomal introns are universal among eukaryotes²⁷, that group II introns originating from the mitochondrion are indeed the most likely precursors of eukaryotic mRNA introns and spliceosomal snRNAs^{16,18,28}, and that many—conceivably most—eukaryotic introns are as old as eukaryotes themselves^{25,26}. More recent are the insights that there is virtually no evolutionary grade detectable in the origin of the spliceosome, which apparently was present in its (almost) fully fledged state in the common ancestor of eukaryotic lineages studied so far²⁷, and that the suspected source of introns—mitochondria, including their anaerobic forms, hydrogenosomes and mitosomes^{20,23,24}—was also present in the common ancestor of contemporary eukaryotes (the only ones whose origin or attributes require explanation).

This suggests that intron origin and spread occurred within a narrow window of evolutionary time: subsequent to the origin of the mitochondrion, but before the diversification of the major eukaryotic lineages. This, in turn, indicates the existence of a turbulent phase of genome evolution in the wake of mitochondrial origin, during which group II introns invaded the host's chromosomes, spread as transposable elements into hundreds—perhaps thousands—of positions that have been conserved to the present, and fragmented into both mRNA introns and snRNA constituents of the spliceosome. Why is this noteworthy?

The organisms that have mitochondria have introns, and a causal link is suspected. But the organisms that have mitochondria and introns also have a cell nucleus, and there are now reasons to suspect a

¹Institute of Botany III, University of Düsseldorf, D-40225 Düsseldorf, Germany. ²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

causal link here as well. The evolutionary relationships of proteins specific to the nuclear envelope and nuclear pore complex reveal that this sizable protein set is a mix of proteins and domains of archaeobacterial and eubacterial origins, along with some eukaryotic innovations, suggesting that the nucleus arose in a cell that already contained a mitochondrial endosymbiont²⁹. Furthermore, gene duplication patterns suggest that the endoplasmic reticulum arose before the nucleus²⁹, as it occurs in the modern cell cycle³⁰. Additionally, evolutionary proteomics show that the nucleolus also contains proteins of both archaeobacterial and eubacterial origins, suggesting that it too arose in a mitochondriate cell³¹. None of the foregoing interpretations needs be correct. However, biologists have traditionally addressed the origin of the nucleus independently of mitochondrial origins³⁰, which need not be correct either. Hence, it is prudent to explore alternative avenues of investigation.

Introns in cytosolic chromosomes

Our salient considerations start with a simple idea proposed by Cavalier-Smith¹⁴: "...the slow splicing of spliceosomal introns would have made it difficult for them to evolve in bacteria (or in mitochondria or chloroplasts) because, if splicing is slow, the coexistence of DNA and functional ribosomes in the same cell compartment would allow ribosomes to translate unspliced premessengers and make incorrect proteins with intron sequences or, if the introns had stop codons, truncated and chimaeric proteins", as similarly expressed by Doolittle³². We require two additional premises for our inference.

As the simplest null hypothesis for the host that acquired the mitochondrion, we assume that it was a prokaryote—not a eukaryote, as previous authors have assumed^{14,30,32}—and, therefore, that it lacked anything similar to a nucleus. In support of that premise, there are prokaryotes (albeit eubacterial) known to harbour eubacterial endosymbionts³³. Thus, in contrast to some current views³⁰, the host need neither have been phagotrophic nor nucleate. Biologists agree that the host was in some way related to archaeobacteria, although the nature of that relationship is still debated³⁴. Either the host had a nucleus, or it did not. Our premise is that it did not; therefore, it was a prokaryote related to archaeobacteria (thus an archaeobacterium), which is compatible with any model for the origin of mitochondria that entails a prokaryotic host^{35,36}. However, our present proposal is independent of the microbial physiology underlying that symbiosis. We also assume that mitochondria stem from α -proteobacteria^{20,23,34–36}. Thus, without specifying a mechanism by which the mitochondrial endosymbiont came to reside within the cytosol of its prokaryotic host, but noting that such entry is possible³³, our inference starts with an α -proteobacterial ancestor of mitochondria and hydrogenosomes^{20,22,23,35} inside an archaeobacterial host^{34–36} with one (or more) cytosolic chromosome(s) (Fig. 1, bottom).

This bipartite cell would not be an immediate success story: it would have nothing but problems instead. Rates of cell division for the host and the symbiont would have to reach comparable levels³⁷. Only progeny that synchronized these rates would persist as consortia. If a host occasionally lyses, symbionts are set free. However, if a symbiont occasionally lyses, a genome's worth of eubacterial DNA is left in the host's cytosol, free to recombine. This situation is similar to Doolittle's ratchet³⁸ but differs in three salient points: the host here is a prokaryote, not a eukaryote; transfer occurs from a persistent resident symbiont, not from undigested meals; and gene transfer from that symbiont is facilitated by the lack of a nuclear membrane. As long as there is more than one symbiont per progeny, symbiont lysis can occur repeatedly, resulting in a constant flow of symbiont DNA into the chromosome(s) of the host.

Intron invasion causes problems

For the host, spreading group II introns that have hitchhiked into its chromosomes become an issue. In modern α -proteobacteria, group II introns occur at up to ~30 copies per genome¹⁸. Prokaryotic group II introns require the reverse transcriptase and maturase activities of their intron-encoded protein for mobility and splicing^{17,18}, have as yet unknown molecular control mechanisms, and are probably kept at bay by purifying selection in large prokaryotic populations^{28,39}. They are usually—but not always—inserted either within intergenic spacers or within mobile elements, like insertion sequences, where they are not really introns but just other mobile elements^{18,28}. Given the antiquity of many eukaryotic intron positions and the observed endonuclease-dependent mobility of modern group II introns in prokaryotes^{17,18}, the initial invasion of eukaryotic introns is most simply envisaged in the form of mobile, *bona fide* group II elements from the mitochondrion—their transition to spliceosome-dependent introns having occurred subsequently at the newly occupied sites. Spreading group II introns constitute a genetic burden, but not

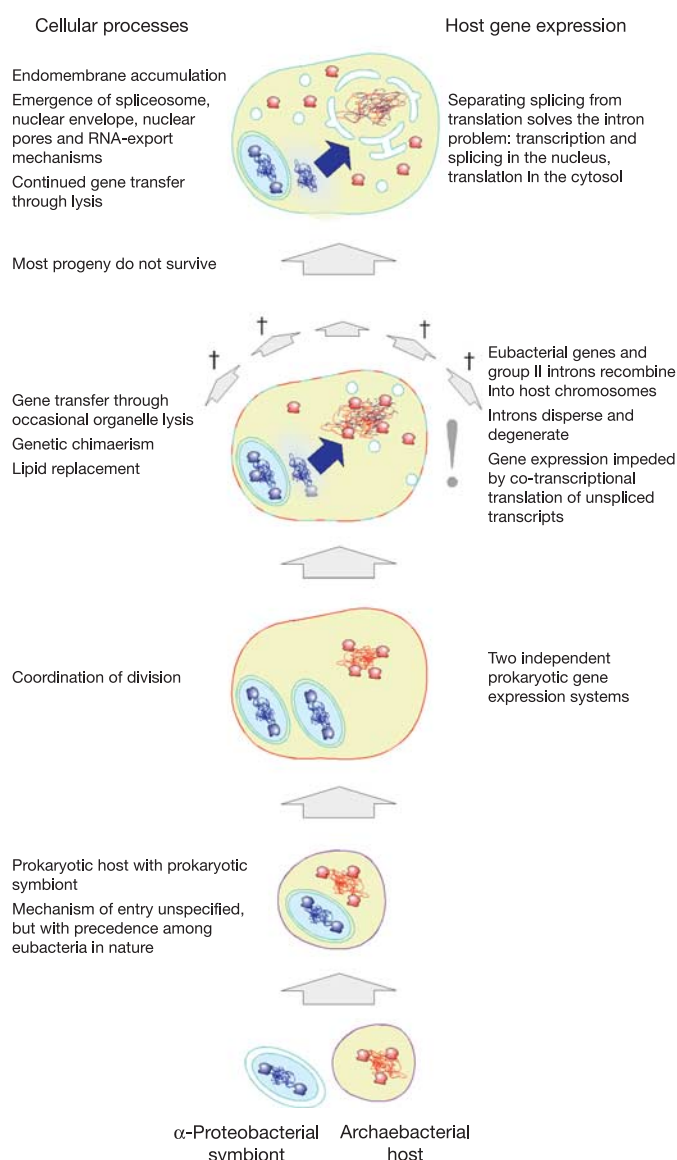


Figure 1 | Origin of nucleus-cytosol compartmentalization in the wake of mitochondrial origin. Blue arrows indicate symbiont-to-host gene transfer. The arrows marked with crosses symbolize the ill fate of most progeny that suffered intron invasion and other endosymbiont-triggered disturbances, resulting in a population bottleneck among progeny from a singular endosymbiotic event. Archaeobacterial and eubacterial features are indicated in red and blue, respectively.

an insurmountable one, because prokaryotes and organelles can effectively express genes that contain group II introns^{18,28}.

A problem of a much more severe nature arises, however, with the mutational decay of group II introns, resulting in inactivation of the maturase and/or RNA structural elements in at least some of the disseminated copies. Modern examples from prokaryotes and organelles suggest that splicing with the help of maturase and RNA structural elements provided by intact group II introns in *trans*¹⁸ could have initially rescued gene expression at such loci, although maturase action in *trans* is much less effective than in *cis*¹⁸. Thus, the decay of the maturase gene in disseminated introns poses a requirement for invention of a new splicing machinery. However, as discussed below, the transition to spliceosome-dependent splicing will also impose an unforgiving demand for inventions in addition to the spliceosome.

Spliceosomes contain five snRNAs and about 200 proteins²⁷, and exist in two distinct forms for removing U2- and U12-dependent introns²⁸. Evidence for an evolutionary transition of group II elements into spliceosomal introns comes from their similar splicing mechanisms^{12,18,28}. Furthermore, the evolutionary conservation of the spliceosome²⁷ and intron positions^{25,26} suggests that this transition occurred in the eukaryote ancestor, entailing the recruitment of group II-derived RNAs in *trans*, the precursors of snRNAs^{18,28}, plus novel accessory proteins²⁷.

The Sm-domain, a protein structural module involved in the still poorly characterized RNA-processing reactions in archaeobacteria⁴⁰, was probably pivotal in that transition. The core complex of today's spliceosomes contains ~20 paralogous Sm-domain proteins that are conserved across eukaryotes⁴⁰. It seems that the protospliceosome recruited the Sm-domain, possibly to replace the maturase, while retaining group II RNA domains (snRNAs) ancestrally germane to the splicing mechanism^{18,28}. While the later evolution of the spliceosome entailed diversification with the recruitment of additional proteins²⁷—leading to greater efficiency—the simpler, ancestral protospliceosome could, in principle, rescue expression of genes containing degenerate group II introns in a maturase-independent manner, but at the dear cost of speed.

Translation in prokaryotes is fast, of the order of 10 amino acids per second⁴¹, whereas splicing by spliceosomes is slow, in the range of 0.005–0.01 intron per second in globins⁴². Although translation in modern eukaryotes is slower, ~1 amino acid per second⁴³, it is still much faster than modern splicing; because our assumed host is a prokaryote, its rate is relevant. The initial, as yet unoptimized protospliceosome—containing the precursors of *bona fide* snRNAs and still recruiting Sm-domain proteins^{27,40}—was surely even slower than the modern one, bringing proper co-transcriptional translation of spliceosome-dependent genes virtually to a halt. Furthermore, in intron-bearing transcripts, the protospliceosome would have had trouble merely gaining access to the ribosome-covered mRNA: a problem that, in modern group II introns, is overcome by the maturase sequestering the ribosome-binding site¹⁸. Thus, ribosomes translating nascent transcripts that bear spliceosomal introns is an extremely unhealthy situation because few functional proteins will ensue^{14,32}, and the prospects of any descendants emerging from this situation are bleak. The only recognizable mechanism operating in favour of this clumsy chimaera is weakened purifying selection operating on its exceptionally small initial population^{28,39}.

Up to this point, the processes that generated introns and the spliceosome require no special inventions or unusual selective pressures; they stem solely from the polarity of the symbiosis, which governs the direction of gene transfer (Fig. 1). Finding a solution to the new problem of slow spliceosomes in the presence of fast and abundant ribosomes required an evolutionary novelty.

Solving the intron problem

There are three obvious routes for solving the problem of possessing spliceosome-dependent introns in co-transcriptionally translated

mRNA. The first solution would be the spontaneous invention of an extremely fast and efficient spliceosome capable of outrunning the ribosomes. This entails a splicing efficiency in the ancestral spliceosome exceeding that of the modern one, and requires the preadaptive evolution of an unselected catalytic function (efficient splicing) in the first protospliceosome before the origin of its mRNA intron substrates, and therefore can be dismissed for that reason. The second is the invention of a mechanism to rapidly and efficiently remove spliceosomal introns from DNA, which—given eukaryote intron antiquity—apparently did not occur. The third solution is the invention of a means to physically separate splicing from translation, allowing the former (slow) process to occur to completion first, before the latter (fast) process sets in. Physical separation in cells usually entails membranes, so the third solution would involve the invention of a membrane separating splicing from translation, with pores sufficiently large and selective enough to export matured ribosomal subunits, mRNA and tRNA. But whence does the nuclear envelope originate?

In the modern cell cycle, the nuclear envelope is continuous with the endoplasmic reticulum⁴⁴ (ER), and their evolutionary origins are probably related^{29,30}. For our present purpose, the existence of a primitive endomembrane consisting of a single eubacterial lipid bilayer is sufficient. The endomembrane system can assume any form, but only sheaths proximal to the host nucleoid will tend to spatiotemporally separate nascent transcripts from ribosomes. Any heritable variation promoting the exclusion of ribosomes from chromosomes until splicing is completed, releasing only processed transcripts to ribosomes in the cytosol, would enhance survival. Progeny that failed to physically separate mRNA processing from translation would not survive, nor would those that failed to invent pore complexes to allow chromosome–cytosol interaction^{29,30}.

The result of our inference is a cell bearing a mitochondrion, chimaeric chromosomes contained within a membrane-bounded splicing compartment (the nucleus), and archaeobacterial ribosomes in the cytosol that translate mature mRNAs from intron-containing and intronless genes of both host and symbiont origin (Fig. 1, top). The invention of the nucleus was mandatory to allow the expression of intron-containing genes in a cell whose ribosomes were faster than its spliceosomes. With the original intron-encoded protein¹⁸ no longer essential, its genome-wide mutational demise would have ended the initial intron invasion and left the ancestral intron set in place, the later emergence of introns in new positions notwithstanding^{25,26}.

A dedicated translation compartment

This view implicates a chromosome-free cytosol, not a nucleus, as the genuinely novel eukaryotic cell compartment: a dedicated translation compartment that is free of transcriptionally active chromosomes. The nucleoplasm is hardly novel in comparison to prokaryotes, because it has retained the job of chromosome maintenance and all basic gene expression functions other than translation. The evolutionary separation of translation from splicing required that mRNA, ribosomal subunits and tRNA mature in the nucleus, culminated by their export—as it occurs in modern eukaryotes, where mRNA export is coupled to splicing, with specific spliceosome components serving as mRNA export factors^{45,46}. This is consistent with our proposal, as is the evidence that the eukaryotic Ntf2 (nuclear transport factor 2; also known as Nutf2) family, to which key proteins of mRNA export—Mex67/TAP/Nxf1 and Mtr2/Nxt1—belong, has homologues in α -proteobacteria²⁹.

In addition to splicing, eukaryotes possess elaborate mRNA surveillance mechanisms, in particular nonsense-mediated decay (NMD), to assure that only correctly processed mature mRNAs are translated, while aberrant mRNAs and those with premature termination codons are degraded^{28,47,48}. The initial intron invasion would have precipitated a requirement for mechanisms to identify

exon junctions and to discriminate exons (with frame) from introns (without frame), as well as properly from improperly spliced transcripts. Thus, NMD might be a direct evolutionary consequence of newly arisen genes-in-pieces. Consistent with that view, the eukaryotic NMD machinery is composed of components derived from both the archaeobacterial translation system and the eubacterial post-segregational cell-killing systems⁴⁹. Thus, even if eukaryotes are found that have no introns left at all, they should possess mRNA surveillance mechanisms^{47,48} as a remnant of their intron-laden past.

Reports that translation occurs to a substantial extent in the mammalian nucleus⁵⁰, the evidence for which remains controversial^{147,48,51,52}, are difficult to reconcile with our proposal. While many nuclear factors are involved in NMD^{47,48}, the participation, if any, of active nuclear protein synthesis in NMD remains uncertain^{47,48}. The need to separate translation from mRNA maturation by the nuclear membrane would seem compelling for all intron-rich eukaryotes, both modern and ancestral.

Conclusion

Our suggestion for the origin of the nucleus differs from previous views on the topic, which either posit that the nuclear membrane was beneficial to (not mandatory for) its inventor by protecting chromosomes from shearing at division³⁰, or offer no plausible selective mechanism at all. The initial eukaryote we infer requires a nucleus only during phases of the cell cycle where genes are expressed, and thus is compatible with either an ancestral or a derived state for closed mitosis⁵³. The archaeobacterial nature of eukaryotic informational genes³⁴ involved in nucleocytoplasmic information storage and expression is readily accommodated by our proposal, as is the α -proteobacterial link with eukaryotic operational (biosynthetic) genes³⁴ of various cytosolic pathways⁵⁴. While contributions from additional symbionts or other gene donors at the origin of the nucleus need not be excluded, our inference requires no more than two partners, each possessing a naturally diverse collection of genes, in order to account for available observations. Neither primitively amitochondriate eukaryotes nor eukaryotic-type spliceosomes in prokaryotes could be accommodated by the premises stated here. As with the spliceosome, we suggest that genes⁵⁵ and protein folds⁵⁶ specific to the eukaryotic lineage arose after the acquisition of mitochondria and before the divergence of the major eukaryotic lineages⁵⁷. An important role for spreading introns in that phase of gene invention, involving rampant exon shuffling¹ mediated by intron sequences that were ancestrally homologous, is implicated.

- Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
- Doolittle, W. F. Genes in pieces: Were they ever together? *Nature* **272**, 581–582 (1978).
- Darnell, J. E. Implications of RNA–RNA splicing in evolution of eukaryotic cells. *Science* **202**, 1257–1260 (1978).
- Doolittle, W. F. Revolutionary concepts in evolutionary cell biology. *Trends Biochem. Sci.* **5**, 147–149 (1980).
- Darnell, J. E. & Doolittle, W. F. Speculations on the early course of evolution. *Proc. Natl Acad. Sci. USA* **83**, 1271–1275 (1986).
- Cech, T. R., Zaug, A. J. & Grabowski, P. J. *In vitro* splicing of the ribosomal-RNA precursor of *Tetrahymena*—involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487–496 (1981).
- Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
- Gilbert, W. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 901–905 (1987).
- Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. & Doolittle, W. F. Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207 (1994).
- Poole, A., Jeffares, D. & Penny, D. Prokaryotes, the new kids on the block. *Bioessays* **21**, 880–889 (1999).
- Forterre, P. & Philippe, H. Where is the root of the universal tree of life? *Bioessays* **21**, 871–879 (1999).
- Jeffares, D. C., Poole, A. M. & Penny, D. Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36 (1998).
- Cech, T. R. The generality of self-splicing RNA: Relationship to nuclear mRNA splicing. *Cell* **44**, 207–210 (1986).

- Cavalier-Smith, T. Intron phylogeny: A new hypothesis. *Trends Genet.* **7**, 145–148 (1991).
- Ferat, J.-L. & Michel, F. Group II self splicing introns in bacteria. *Nature* **364**, 358–361 (1993).
- Roger, A. J. & Doolittle, W. F. Why introns-in-pieces? *Nature* **364**, 289–290 (1993).
- Cousineau, B., Lawrence, S., Smith, D. & Belfort, M. Retrotransposition of a bacterial group II intron. *Nature* **404**, 1018–1021 (2000); correction **414**, 84 (2001).
- Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annu. Rev. Genet.* **38**, 1–35 (2004).
- Nixon, J. E. *et al.* A spliceosomal intron in *Giardia lamblia*. *Proc. Natl Acad. Sci. USA* **99**, 3701–3705 (2002).
- Tovar, J. *et al.* Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* **426**, 172–176 (2003).
- Simpson, A. G., MacQuarrie, E. K. & Roger, A. J. Early origin of canonical introns. *Nature* **419**, 270 (2002).
- Vanacova, S., Yan, W., Carlton, J. M. & Johnson, P. J. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc. Natl Acad. Sci. USA* **102**, 4430–4435 (2005).
- Hrdy, I. *et al.* *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**, 618–622 (2004).
- van der Giezen, M. & Tovar, J. Degenerate mitochondria. *EMBO Rep.* **6**, 525–530 (2005).
- Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13**, 1512–1517 (2003).
- Roy, S. W. & Gilbert, W. Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proc. Natl Acad. Sci. USA* **102**, 5773–5778 (2005).
- Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**, 1053–1066 (2005).
- Lynch, M. & Richardson, A. O. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **12**, 701–710 (2002).
- Mans, B. J., Anantharaman, V., Aravind, L. & Koonin, E. V. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* **3**, 1612–1637 (2004).
- Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* **52**, 297–354 (2002).
- Staub, E., Fizev, P., Rosenthal, A. & Hinzmann, B. Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. *Bioessays* **26**, 567–581 (2004).
- Doolittle, W. F. The origin of introns. *Curr. Biol.* **1**, 145–146 (1991).
- von Dohlen, C. D., Kohler, S., Alsop, S. T. & McManus, W. R. Realybug β -proteobacterial symbionts contain γ -proteobacterial symbionts. *Nature* **412**, 433–436 (2001).
- Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).
- Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
- Vellai, T., Takács, K. & Vida, G. A new aspect on the origin and evolution of eukaryotes. *J. Mol. Evol.* **46**, 499–507 (1998).
- Finlay, B. J. & Fenchel, T. An anaerobic ciliate as a natural chemostat for the growth of endosymbiotic methanogens. *Eur. J. Protistol.* **28**, 127–137 (1992).
- Doolittle, W. F. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
- Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
- Khusial, P., Plaag, R. & Zieve, G. W. LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends Biochem. Sci.* **30**, 522–528 (2005).
- Sorensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in *E. coli*. *J. Mol. Biol.* **207**, 365–377 (1989).
- Audibert, A., Weil, D. & Dautry, F. *In vivo* kinetics of mRNA splicing and transport in mammalian cells. *Mol. Cell. Biol.* **22**, 6706–6718 (2002).
- Palmiter, R. D. Quantitation of parameters that determine the rate of ovalbumin synthesis. *Cell* **4**, 189–197 (1975).
- Levine, T. & Rabouille, C. Endoplasmic reticulum: One continuous network compartmentalized by extrinsic cues. *Curr. Opin. Cell Biol.* **17**, 362–368 (2005).
- Reed, R. & Hurt, E. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* **108**, 523–531 (2002).
- Luo, M. L. *et al.* Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. *Nature* **413**, 644–647 (2001).
- Fasken, M. B. & Corbett, A. H. Process or perish: Quality control in mRNA biogenesis. *Nature Struct. Mol. Biol.* **6**, 482–488 (2005).
- Maquat, L. E. Nonsense mediated mRNA decay: Splicing, translation and mRNP dynamics. *Nature Rev. Mol. Cell Biol.* **5**, 89–99 (2004).
- Anantharaman, V. & Aravind, L. New connections in the prokaryotic toxin–antitoxin network: Relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol.* **4**, R81 (2003).

50. Iborra, F. J., Jackson, D. A. & Cook, P. R. Coupled transcription and translation within nuclei of mammalian cells. *Science* **293**, 1139–1142 (2001).
51. Dahlberg, J. E. & Lund, E. Does protein synthesis occur in the nucleus? *Curr. Opin. Cell Biol.* **16**, 335–338 (2004).
52. Cosson, B. & Philippe, M. Looking for nuclear translation using *Xenopus* oocytes. *Biol. Cell.* **95**, 321–325 (2003).
53. Heath, I. B. Variant mitoses in lower eukaryotes: Indicators of the evolution of mitosis? *Int. Rev. Cytol.* **64**, 1–80 (1980).
54. Gabaldón, T. & Huynen, M. A. Reconstruction of the proto-mitochondrial metabolism. *Science* **301**, 609 (2003).
55. Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G. & Koonin, E. V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **33**, 4626–4638 (2005).
56. Yang, S., Doolittle, R. F. & Bourne, P. E. Phylogeny determined by protein domain content. *Proc. Natl Acad. Sci. USA* **102**, 373–378 (2005).
57. Adl, S. M. et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52**, 399–451 (2005).

Acknowledgements We thank the DFG (W.M.) and the NIH intramural research program (E.V.K.) for financial support, and M. Embley, T. Dagan, I. Rogozin, T. Senkevich and Y. Wolf for discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to W.M. (w.martin@uni-duesseldorf.de).

ARTICLES

Plate-boundary deformation associated with the great Sumatra–Andaman earthquake

Cecep Subarya¹, Mohamed Chlieh², Linette Prawirodirdjo³, Jean-Philippe Avouac², Yehuda Bock³, Kerry Sieh², Aron J. Meltzner², Danny H. Natawidjaja⁴ & Robert McCaffrey⁵

The Sumatra–Andaman earthquake of 26 December 2004 is the first giant earthquake (moment magnitude $M_w > 9.0$) to have occurred since the advent of modern space-based geodesy and broadband seismology. It therefore provides an unprecedented opportunity to investigate the characteristics of one of these enormous and rare events. Here we report estimates of the ground displacement associated with this event, using near-field Global Positioning System (GPS) surveys in northwestern Sumatra combined with *in situ* and remote observations of the vertical motion of coral reefs. These data show that the earthquake was generated by rupture of the Sunda subduction megathrust over a distance of $>1,500$ kilometres and a width of <150 kilometres. Megathrust slip exceeded 20 metres offshore northern Sumatra, mostly at depths shallower than 30 kilometres. Comparison of the geodetically and seismically inferred slip distribution indicates that ~ 30 per cent additional fault slip accrued in the 1.5 months following the 500-second-long seismic rupture. Both seismic and aseismic slip before our re-occupation of GPS sites occurred on the shallow portion of the megathrust, where the large Aceh tsunami originated. Slip tapers off abruptly along strike beneath Simeulue Island at the southeastern edge of the rupture, where the earthquake nucleated and where an $M_w = 7.2$ earthquake occurred in late 2002. This edge also abuts the northern limit of slip in the 28 March 2005 $M_w = 8.7$ Nias–Simeulue earthquake.

The great Sumatra–Andaman earthquake of 2004 was produced by rupture of the Sunda subduction megathrust, along which the Indian and Australian plates subduct northeastward beneath the Sunda shelf (Fig. 1). Southeast of Sumatra, at Java, convergence is nearly orthogonal to the plate boundary at ~ 63 – 68 mm yr⁻¹ (refs 1, 2). Along Sumatra the convergence is oblique to the trench and the relative plate motion is partitioned into nearly perpendicular thrusting on the megathrust at ~ 45 mm yr⁻¹ and trench-parallel, right-lateral slip along the Sumatra fault at ~ 11 to 28 mm yr⁻¹ (refs 3, 4). The convergence rate normal to the trench is ~ 40 mm yr⁻¹ near the 2004 epicentre off northern Sumatra and decreases northwards as the megathrust strike becomes nearly parallel to the direction of relative plate motion. North of 8° N, sparse geodetic data suggest a convergence rate normal to the trench of between 14 and 34 mm yr⁻¹ (refs 5, 6).

The Sumatran section of the Sunda megathrust generated great earthquakes south of the 2004 event in 1797, 1833 and 1861 (refs 7–9) but there is no historical record of giant earthquakes to the north, between Sumatra and Myanmar (Fig. 1).

Analyses of high-frequency seismic records of the December 2004 earthquake obtained from the Global Seismic Network¹⁰, from an array of seismic stations in Thailand¹¹ and from T-waves recorded in the Indian Ocean^{11,12}, indicate that the rupture took about 500 s to propagate a straight-line distance of $\sim 1,300$ km from the hypocentre in northern Sumatra to the northern Andaman Islands. This rupture area roughly coincides with the distribution of aftershocks^{6,11} (Fig. 1). A model of the slip history and its spatial distribution obtained by

combining body waves and surface waves yielded a total seismic moment for the earthquake of 6.5×10^{22} N m, released mostly between latitudes 2° N and 10° N, corresponding to $M_w = 9.1$ (refs 13, 14).

We report here on near-field GPS observations of deformation and *in situ* and remotely sensed observations of uplift and subsidence of coral reefs. We use these to constrain the distribution of slip on the Sunda megathrust during and soon after the 26 December 2004 earthquake, and to compare it to slip models derived from seismic data. These geodetic data allow us to model in great detail the slip distribution west of northern Sumatra, the region of greatest devastation.

GPS measurements

Our re-survey of GPS monuments in northern Sumatra between 28 January and 19 February 2005 reveals combined coseismic and postseismic displacements of up to several metres associated with the earthquake (Fig. 2). One set of monuments was surveyed three or four times in the years 1991–2001 (refs 15 and 16) and provides a record of pre-earthquake interseismic velocities¹. This GPS network includes lines of closely spaced points across the Sumatran fault⁴, including one in northernmost Aceh, across the region of greatest devastation from the ensuing tsunami. A second set of monuments was surveyed only once before the earthquake by the Indonesian National Coordinating Agency for Surveys and Mapping (BAKOSURTANAL) as part of geodetic control for Sumatra. One of these sites (R171) is on Selaut Besar, a small island north of Simeulue

¹National Coordinating Agency for Surveys and Mapping, Cibinong 16911, Indonesia. ²Tectonics Observatory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. ³Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA. ⁴Research Center for Geotechnology, Indonesian Institute of Sciences, Bandung 40135, Indonesia. ⁵Department of Earth and Environmental Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180, USA.

Island, only about 50 km from the earthquake's epicentre (Fig. 3a). We used the decade-long pre-earthquake GPS measurements to construct a kinematic model of interseismic deformation that allowed us to correct the measured displacements for steady interseismic motions (Table 2 in Supplementary Information 1).

Continuous GPS data from BAKOSURTANAL's site SAMP (Fig. 3a) reveal a clear record of coseismic and postseismic deformation. The daily time series shows a coseismic horizontal displacement of 138 mm that increased logarithmically with time after the main shock by $\sim 15\%$ over 15 days, and $\sim 25\%$ over 30 days. For comparison continuous measurements at site PHKT on the island of Phuket indicate a coseismic slip of 270 mm, which increased by $\sim 22\%$ over 30 days (ref. 17). Although these two records reveal significant post-earthquake motion, they do not show how widespread or variable it was. Continuous GPS data from the Sumatran GPS Array (SuGAR; <http://www.tectonics.caltech.edu/sumatra/data.html>), more than 300 km south of the epicentre, show coseismic displacements typically less than 10 mm and no detectable postseismic transients.

We processed the raw survey-mode and continuous GPS data with the GAMIT/GLOBK software (<http://www.gpsg.mit.edu/~simon/gtgk/>)¹⁸. The data were analysed in 24-h segments (0–24 h GMT) with data from ten additional continuous GPS sites on Java, the

Cocos Islands, Diego Garcia, Singapore, India, Australia and Guam. These solutions were combined with global GPS network solutions produced routinely at the Scripps Orbit and Permanent Array Center (<http://sopac.ucsd.edu>) to determine the GPS velocities and displacements and their uncertainties with respect to the ITRF2000 reference frame¹⁹.

Uplift determined from field measurement of coral heads

At the southern end of the rupture, coral heads enabled measurement of uplift. We used their 'micro-atoll' morphology to measure pre- and post-earthquake sea level^{9,20,21}. We measured these sea level proxies on 17 and 18 January and on 5 February 2005 at ten locations around Simeulue Island. We found the pre-quake highest level of survival to be systematically 0.2 to 1.5 m higher than the post-quake level, with values rising towards the northwest (Fig. 2 and Supplementary Information 1). Differences in low tide levels before and after the earthquake, computed according to an ocean-tide model²², lead to adjustments of just a few centimetres²³. The accuracy of the measurements, about ± 50 to 100 mm, is only ~ 2 –3 times worse than the vertical accuracy of typical field GPS geodetic measurements. The advantage of the coral measurements is that they form a dense array of points that constrains the tilting of Simeulue and therefore the gradient in slip at the southern end of the underlying megathrust. We also collected less quantitative evidence of submergence around the southern half of Simeulue Island; at two localities these are eyewitness accounts of sea level changes. At another place we measured the depth of flooding of a well-drained locality where residents said water had never stood before.

Determination of uplift and subsidence from remote sensing

We used satellite imagery (ASTER, SPOT, IKONOS, QUICKBIRD and LANDSAT) to assess changes in relative sea level associated with the earthquake²³. Because the colour and brightness of a reef in an image depend on water depth above the reef, changes in water depths of several centimetres or more are recognizable on the images. We examined satellite images of the Andaman and Nicobar islands and northwestern Sumatra to identify areas where reef or land exposure changed following the earthquake.

Satellite images acquired before 26 December 2004 were compared with images acquired between 26 December 2004 and 28 March 2005. We used a tidal model²² ($1\text{-}\sigma$ uncertainty of ~ 5 cm) to determine the relative sea surface height at each location at the acquisition time of each image. To document the uplift of a reef, we looked for a post-earthquake image with more reef exposure than a pre-earthquake image of the same area taken at a lower tide; in that case, the difference in sea surface height between the two images provides a minimum amount of uplift. Similarly, a pre-earthquake image with more exposure than a post-earthquake image at a lower tide indicates subsidence; in this case, the difference in sea surface height gives the minimum subsidence.

Although we can provide both maximum and minimum constraints on uplift or subsidence in a few locations, in most cases this method is limited by the tidal range; where uplift or subsidence exceeded the tidal range, we can provide only a minimum bound on the amount of tectonic elevation change. Nonetheless, the extrema of vertical displacements and the sign of the elevation change at a location are robust. Altogether we made such observations at 156 locations (Fig. 2). These data show detectable uplift from Simeulue to Preparis Island (Myanmar) over a distance of 1,600 km along the trench²³.

Fault slip distribution from inversion of geodetic data

Vertical ground displacements determined from the various techniques show a characteristic pattern: a region of uplift nearer the trench and a region of subsidence away from it (Fig. 2). The pivot line, which separates the areas of coseismic uplift and subsidence, approximates the easternmost extent of slip on the fault surface

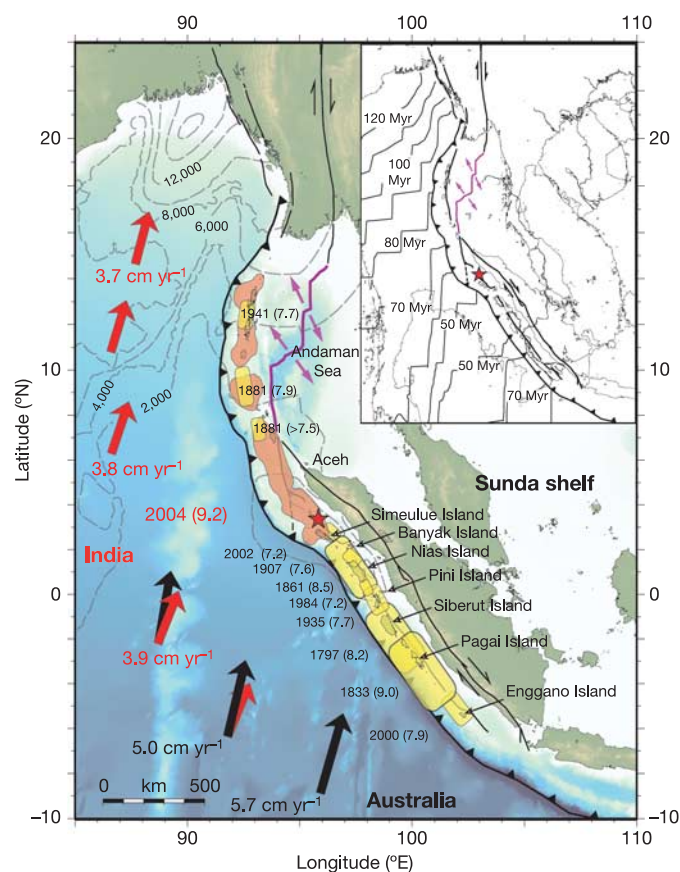


Figure 1 | Tectonic setting and ruptures of major interplate earthquakes along the Sunda megathrust. The yellow patches are estimated rupture areas of known large subduction events between 1797 and 2004 (refs 7, 9, 20). Orange patches depict the 2004 Sumatra–Andaman rupture where slip was 5 m or more. Tectonic features are simplified from Curray⁴³ and Natawidjaja *et al.*²⁰. The boundary between Australia and India is a diffuse plate boundary between $\sim 5^\circ$ S and 8° N (ref. 44). Plate velocities of Australia (black arrows) and India (red arrows) relative to Sunda were computed from a regional kinematic model¹. Dashed lines are contours of sediment thickness at intervals of 2,000 m. The inset shows that the age of the sea floor increases northwards, from ~ 50 Myr in the epicentral area to 80–120 Myr at the latitude of the Andaman islands.

below. Where constrained by both uplift and subsidence observations, the pivot line of the 2004 earthquake lies between 80 and 120 km from the trench. In the area of the Nicobar Islands, all of which subsided, the pivot line is close to the westernmost islands²³, less than 150 km from the trench. Observations of both uplift on north-western Simeulue Island and subsidence on the southernmost part of the island also indicate that the southern, lateral limit of slip is beneath the island. These simple observations imply that the rupture area was confined to the shallow part of the subduction zone within about 150 km of the trench and did not extend south of Simeulue Island.

We estimate the three-dimensional distribution of slip on the megathrust by inverting the geodetic observations described above, GPS measurements from the Nicobar and Andaman islands (<http://www.seires.net/content/view/123/52/>, CESS website)²⁴, and continuous GPS offsets in Phuket and Medan (Fig. 3a, b). Because the uplift and survey-mode GPS observations were made a month or so after the earthquake, they probably contain displacements due to aftershocks and postseismic slip. Displacements directly associated with aftershocks are, however, relatively minor, because the total seismic moment from aftershocks is less than one per cent of that of the mainshock. We followed a two-step procedure in the inversion of the geodetic data (details in the Supplementary Information). To facilitate direct comparison of seismic and geodetic slip models, we first inverted the geodetic data using the same simplified fault geometry: three planar faults, and layered structure as in previous seismological models¹⁴ (model A, Fig. 3a). Then, to assess the sensitivity of the results to the fault geometry and seismic velocity structure, we used a more realistic three-dimensional fault geometry, in a homogeneous half-space (model B, Fig. 3b). Details on these two models are given in the Supplementary Information.

In model A, the fault is represented by three overlapping planar segments with different strikes, and dip angles increasing from 12° in the south to 17.5° in the north, with the slab extending to about 125 km depth. In this model, our best estimate of the geodetic moment is 8.8×10^{22} N m, corresponding to a magnitude of $M_w = 9.22$. The weighted root-mean-square is 1.8 cm, corresponding to a reduced χ^2 of 2.44. Sensitivity and resolution tests suggest that the model probably provides a lower bound on the estimated moment required to fit the geodetic data. The scalar seismic moment for the best-fit geodetic model is ~30% greater than the seismological estimate (Fig. 4). The geodetic model predicts remarkably well the azimuths of coseismic displacements observed at continuous GPS stations in Thailand and Malaysia¹⁷ (Fig. 3a), but the amplitudes are systematically larger by an average of 26%. The seismological model¹⁴ predicts displacements which also agree well with these azimuths, but underpredicts the amplitudes. When only the coseismic data (representing geodetic displacements over one day) from Thailand and Malaysia and at Medan are inverted, the geodetic moment is constrained to about 6.8×10^{22} N m (as also estimated by Vigny *et al.*¹⁷), which is close to the seismic estimate. We conclude that the excess moment of the geodetic model over the seismic model, equivalent to about a $M_w = 8.7$ earthquake, reflects aseismic afterslip in the weeks following the earthquake, rather than slow aseismic slip during the first day after the earthquake, as proposed in some early studies^{6,13}.

In model B, the slip distribution is represented as three-parameter gaussian functions of depth along 26 trench-normal profiles between 1° N and 16° N. In cross-section, the modelled fault geometry, based largely on earthquake distributions, is curved downward and is on average steeper in the northern profiles. The 78 free parameters are estimated by least-squares fit to the 287 weighted observations,

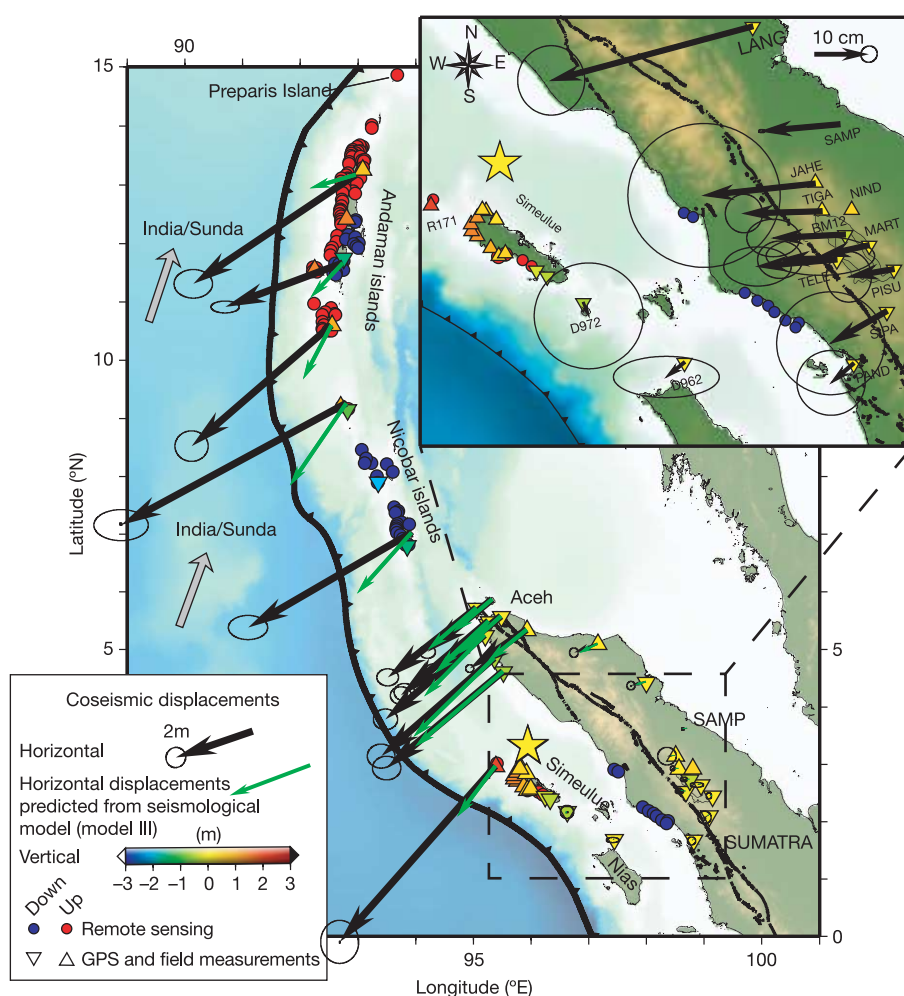


Figure 2 | Comparison of near-field geodetic measurements (black arrows) with predictions (green arrows) of the seismic model III of ref. 14. This comparison suggests that the geodetic data require more slip, a different spatial distribution of slip, or both. The inset shows a close-up of the Simeulue area. Horizontal displacements are shown with 95% confidence ellipses (see tables in Supplementary Information 1). Vectors in the Andaman and Nicobar islands are from CESS (<http://www.seires.net/content/view/123/52/>).

giving a reduced χ^2 of 0.83, indicating the observations are matched closely at their levels of uncertainty. It also yields a seismic moment of $M_w = 9.22$ and slip that varies markedly along strike. Both models show three distinct patches of high slip from 4° to 6° N, 8° to 10° N, and 12° to 13.75° N. These patches probably correspond to the three distinct bursts of energy seen in the seismological inversions and attributed to patches of high slip¹⁴. Both models suggest a minimum rupture length of about 1,400 km, based on the area within which slip exceeded 5 m. Given the uplift documented at Preparis Island²³ the rupture must have been somewhat longer: about 1,600 km. Both models display a prominent trough in slip values from about 7° to 8° N, which may reflect a lack of local geodetic measurements. Slip near the epicentre was relatively low (<15 m) but ramped up

dramatically northward to >20 m. Model A places all slip at depths shallower than 50 km. The more realistic curved geometry of model B yields a shallower slip distribution in which most slip was shallower than 25 km depth. This shallowness of slip is the principal reason that the rupture generated the great tsunami.

Discussion

The 2004 Sumatra–Andaman earthquake illuminates the rupture processes of giant earthquakes. Such earthquakes are so rare that we have relied largely on empirical correlations between properties of megathrust earthquakes and their subduction zones to understand them. One widely accepted relationship^{25,26} is that maximum earthquake magnitude on a given thrust increases linearly with convergence

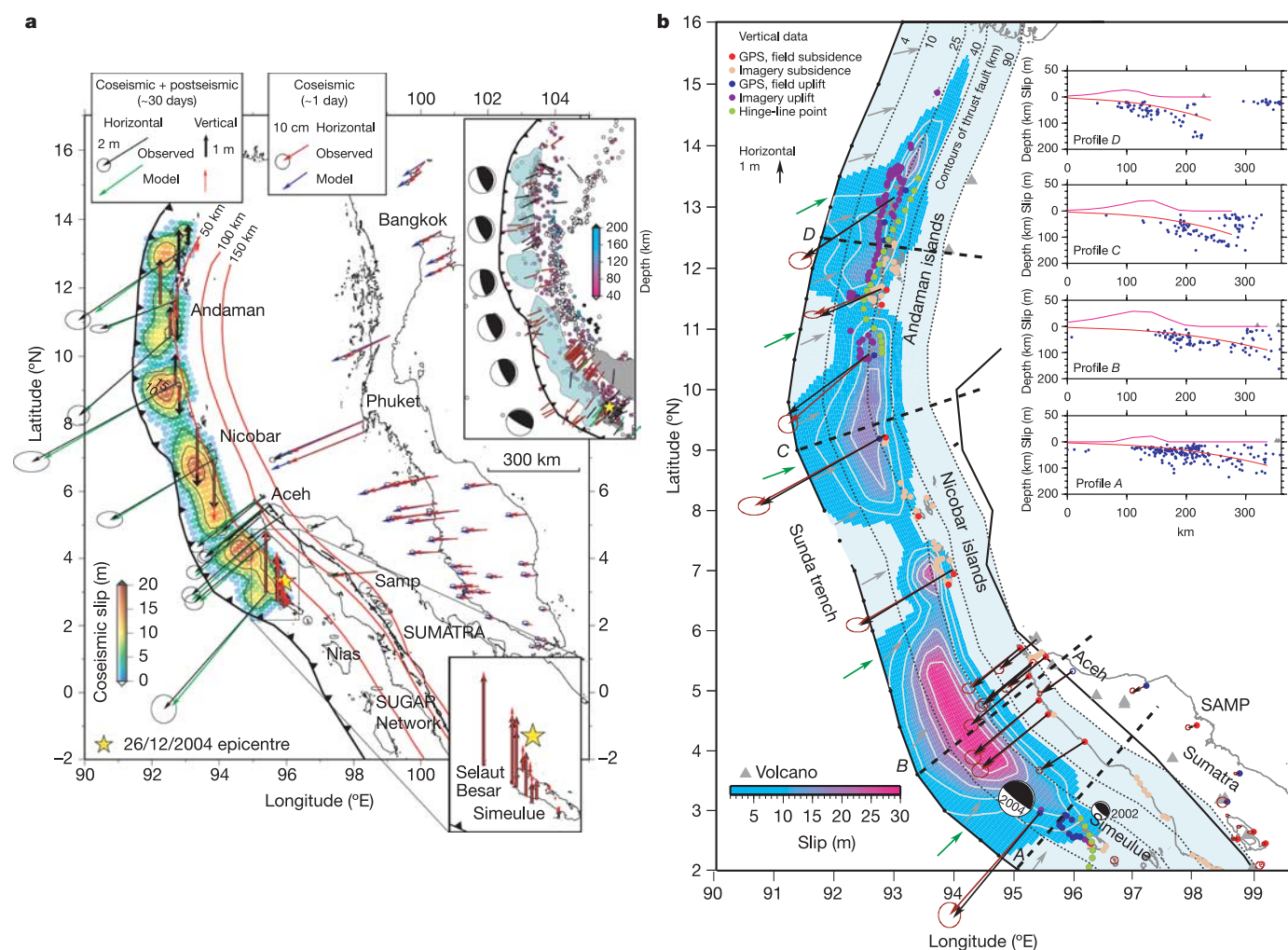


Figure 3 | Fault slip distribution determined from the geodetic data.

a, Model A. The distribution of combined coseismic and one-month postseismic slip on the Sunda megathrust estimated from inversion of geodetic data shown in Fig. 2, including 30-day estimates of displacements from the permanent GPS stations at Medan (SAMP) and Phuket (PHKT)¹⁷. Black contour lines of slip are at 5-m intervals. Displacements computed from this model (green arrows for the horizontal and red arrows for the vertical at the sites used in the inversion, blue arrows at the other sites) are compared with the survey-mode observations and with displacements over the first day estimated at continuous GPS stations in Thailand and Malaysia¹⁷ (red arrows). Displacements are shown with 95% confidence ellipses. The comparison shows that significant postseismic displacements occurred in the first month following the rupture. The lower inset in **a** is a close-up view of predicted and measured vertical displacements on Simeulue Island. The upper inset in **a** shows a comparison of ruptured area (where slip exceeds 5 m in the model) with seismicity before the Sumatra–Andaman earthquake (1964–2002)⁴⁴. ‘Beach ball’ symbols show local mechanisms determined

from the geodetic model. Each focal mechanism corresponds to the summation of moment tensor within a 2° wide latitudinal bin. Slip vector azimuth of aftershocks (red lines) and foreshocks (black lines) are nearly parallel to slip azimuth during the main shock. **b**, Model B. The distribution of co- and post-seismic seismic slip on the Sunda megathrust estimated from inversion of geodetic data. Light contours of slip are at 5-m intervals starting at 5 m. Red vectors (with 95% confidence ellipses) are observed displacements and black are predicted. Coloured dots show locations of uplift constraints; those that are outlined were not fitted at the $2\text{-}\sigma$ level (hinge-line points were not used in this inversion). Small arrows near the Sunda trench show seismological estimates of coseismic slip directions in green⁴⁵ and geodetic estimates in grey. The insets show trench-normal profiles of earthquakes⁴⁴ (blue dots), megathrust (red curve), and slip amplitudes (purple curves). This model indicates up to 30 m of slip at depths of only 12–20 km, northwest of nucleation and where the large Aceh tsunami originated.

rate and decreases linearly with subducting plate age. The relatively small Car Nicobar and Andaman earthquakes of 1881 and 1941 (ref. 6), enveloped within the northern part of the 2004 rupture, fitted this pattern—the subducting lithosphere is old and converging at a moderate rate (Fig. 1). Backarc extension, like that east of the northern part of the 2004 rupture, has also been associated with low-magnitude maximum magnitudes^{27,28}. The size of the 2004 earthquake is clearly at odds with these concepts. Thus, these empirical relationships must neglect important physical processes governing the seismicity of subduction zones. An alternative explanation for the distribution of large earthquakes at subduction zones over the past century is that these correlations result from a sample period that is too short—faster slipping subduction zones will on average produce larger earthquakes in a given time period because the repeat time of an earthquake of a given size is inversely related to the fault slip rate²⁹. Perhaps we must now consider the possibility that, given a suitable length of time, any megathrust fault can produce an earthquake whose size is limited only by the available area of the locked fault plane. From 0.5° to about 6°S, the Sunda megathrust has produced two giant earthquakes in recorded history^{7,30}, and geodetic measurements show it has been fully locked above a depth of 40 to 55 km for at least the past 50 years^{1,31}. Further south, the long section of the Sunda megathrust adjacent to the densely populated island of Java subducts very old lithosphere and should, according to previous wisdom, not produce great quakes. The degree of locking of the megathrust and possibility of great earthquakes there should now be investigated.

Could we have forecast the width of the 2004 megathrust rupture? Analyses of the geodetic and paleogeodetic records of interseismic deformation a few hundred kilometres south of the epicentre suggest that the depth of the downdip end of the locked zone varies from ~30 to 55 km (refs 15, 31–33). No published geodetic or paleogeodetic data were available to constrain interseismic deformation in the area of the 2004 rupture. However, background seismicity provided a clue. Along intracontinental megathrusts, background seismicity tends to cluster around the downdip end of the locked fault zone³⁴. Before December 2004, seismicity was clustered near the downdip end of the future rupture zone, at depths between 40 and 50 km (Fig. 3a). This suggests that the rupture remained confined to the shallow portion of the fault zone that was locked before the great event. Our results thus support the use of background seismicity as one indicator of the down-dip limits of future seismic ruptures.

Substantial afterslip followed the 2004 coseismic rupture. The geodetic data suggest that slip, equivalent to an $M_w = 8.7$ earthquake, occurred along the plate interface in the month following the 2004 earthquake. Afterslip downdip of the coseismic rupture is not uncommon^{35–38}, but the data do not reveal significant deep slip. The

correlation between the zone with high slip determined from geodesy and the area that generated high-frequency body waves¹⁴ suggests that early afterslip occurred on or close to the fault patch that underwent coseismic slip. It is possible that a significant fraction of afterslip occurred updip of the seismically ruptured area. However, the details of the spatial and temporal evolution of slip on the shallow plate interface during and after the event cannot be constrained because of the lack of data close to the trench.

The great horizontal extent of the rupture, which ultimately led to the great magnitude, would have been far more difficult to forecast. The along-strike variability of the coseismic slip distribution (Fig. 3) might reflect past earthquake history, with areas of low slip corresponding to patches that ruptured during past events, or could indicate that the megathrust fault plane is a mix of aseismically slipping areas characterized by a rate-strengthening friction law, and areas of stick-slip behaviour, characterized by a rate-weakening friction law. The latter hypothesis is more plausible, given the correlation of historical $M_w > 7$ earthquakes with high-slip patches of 2004. The frictional and seismic properties of fault zones are thought to depend on a number of factors, including lithology, temperature, pore pressure and normal stress³⁹, that could act jointly to produce variable behaviour. Because aseismic creep is thermally activated, temperature might limit the bottom of the locked fault zone by promoting aseismic slip at depth⁴⁰. Another possibility is that the downdip end of the locked fault zone coincides with the intersection of the plate interface and the forearc Moho, because stable sliding slip could occur along the serpentinized mantle wedge⁴⁰. In the Sumatra–Andaman case, we discount this possibility because the forearc Moho intersects the megathrust well updip of the bottom of the interseismic locked zone³³. To assess whether or not temperature might control the downdip extent of the ruptured area, we estimate the along-strike depth of the 350 °C isotherm, a commonly assumed temperature at the downdip end of the locked section of subduction megathrusts. For an average shear stress between 20 to 40 MPa on the fault and for the variety of subduction dip-angles, this depth is around 40 km in the epicentral area and does not vary much along strike from northern Sumatra to the northern Andaman Islands. This near-constancy in fault zone temperature occurs because the lower heat flow at the top of the older lithosphere in the north has longer to transmit heat to the upper plate owing to the lower trench-normal slip rate.

Still, temperature cannot easily explain short-wavelength lateral variations of frictional properties; other factors must control changes in behaviour. The high proportion of aseismic slip on the 2004 rupture plane may, for example, be due to a lubricating or pore-pressure effect of sediments from the Bengal fan subducting down along the megathrust. The thickness of the sediment reaching the trench is indeed great along the entire rupture, decreasing gradually southwards from more than 4 km to about 1 km (Fig. 1).

The large proportion of afterslip on the 2004 rupture and the irregular coseismic slip pattern might indicate that much of the megathrust slips aseismically. If the proportion of aseismic to seismic slip during the 2004 Sumatra–Andaman earthquake is representative of the long-term average, aseismic slip might be of the order of 0.5 or greater in the Andaman and Nicobar area. The large fraction of aseismic slip may account for the common observation that seismic moment release along subduction zones falls short of the value estimated from the long-term slip rate along the seismogenic portion of the plate interface^{41,42}.

We estimate a nominal repeat time for the 2004 event by dividing the quake's potency (slip times rupture area; $1.7 \pm 0.1 \times 10^{12} \text{ m}^3$) by the long-term potency rate ($3\text{--}7 \times 10^9 \text{ m}^3 \text{ yr}^{-1}$), estimated from the area of the subduction interface north of 2°N (about $2.0 \times 10^5 \text{ km}^2$), and the long-term average slip rate ($24 \pm 10 \text{ mm yr}^{-1}$). If all this slip was released only by the repetition of events like the Sumatra–Andaman earthquake, such events would occur on average every 230–600 years; if half of the slip is aseismic, or taken up by smaller

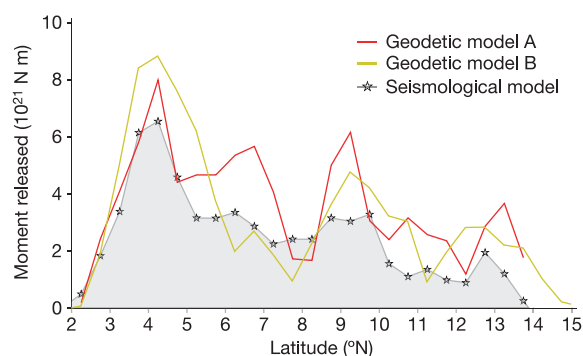


Figure 4 | Latitudinal variations of scalar moments as determined from seismic waveforms (model III of ref. 14) and from geodetic data. Moment released per half degree in latitude. Both geodetic models imply a rougher slip distribution than the seismic model. The total moment for geodetic model A ($8.78 \times 10^{22} \text{ N m}$) exceeds the seismic moment by $30\% \pm 12\%$. This excess presumably reflects afterslip during the 30 days following the main shock.

events such as the events in 1881 or 1941, the recurrence time would double. Such long average return periods are consistent with no historical record of prior events.

A striking feature of the slip distributions we derived is the abrupt southern termination, required by the rapid southward decrease in coral uplift. Our measurements show that the southern limit of uplift in 2004 is approximately coincident with the northern limit of uplift during the 28 March 2005, $M_w = 8.7$ Nias–Simeulue earthquake (<http://www.gps.caltech.edu/~jichen/Earthquake/2005/sumatra/preliminary/sumatra.html>). The proximity of the 2004 and 2005 uplift terminations and a $M_w = 7.2$ foreshock on 2 November 2002 could reflect the presence of a structural feature that is an impediment to rupture propagation. Perhaps the long north-trending fracture zone on the seafloor of the Indian plate that projects to this point⁷ has created a structural or rheological complexity in the megathrust beneath central Simeulue Island. Similar structural discontinuities on the sea floor may have influenced the termination points of large megathrust ruptures in 1861, 1833 and 1935 (refs 7, 20), but the exact mechanism by which they might have done so remains elusive.

Received 20 July; accepted 9 December 2005.

- Bock, Y. *et al.* Crustal motion in Indonesia from Global Positioning System measurements. *J. Geophys. Res.* **108**(B8), 2367, doi:10.1029/2001JB000324 (2003).
- Michel, G. *et al.* Crustal motion and block behaviour in SE-Asia from GPS measurements. *Earth Phys. Sci. Lett.* **187**, 239–244 (2001).
- Sieh, K. & Natawidjaja, D. Neotectonics of the Sumatran fault, Indonesia. *J. Geophys. Res.* **105**, 28295–28326 (2000).
- Genrich, J. F. *et al.* Distribution of slip at the northern Sumatran fault system. *J. Geophys. Res.* **105**, 722–743 (2000).
- Paul, J. *et al.* The motion and active deformation of India. *Geophys. Res. Lett.* **28**, 647–650 (2001).
- Bilham, R., Engdahl, R., Feldl, N. & Satyabala, S. P. Partial and complete rupture of the Indo-Andaman plate boundary 1847–2004. *Seismol. Res. Lett.* **76**, 299–311 (2005).
- Newcomb, K. R. & McCann, W. R. Seismic history and seismotectonics of the Sunda Arc. *J. Geophys. Res.* **92**, 421–439 (1987).
- Zachariassen, J., Sieh, K., Taylor, F. W. R., Edwards, R. L. & Hantoro, W. S. Submergence and uplift associated with the giant 1833 Sumatran subduction earthquake: Evidence from coral microatolls. *J. Geophys. Res.* **104**(B1), 895–920 (1999).
- Sieh, K., Natawidjaja, D., Chlieh, M., Galetzka, J. & Avouac, J. P. The giant subduction earthquakes of 1797 and 1833, West Sumatra: Characteristic couplets, uncharacteristic slip. *Eos* **85** (Fall Meeting Suppl.), abstr. T12B–04 (2004).
- Ni, S., Kanamori, H. & Helmlinger, D. Energy radiation from the Sumatra earthquake. *Nature* **434**, 582 (2005).
- Guilbert, J., Roueff, A., Vergoz, J. & Cansi, Y. An original image of the seismic rupture of the Sumatra $M_w = 9.0$ using PMCC from seismic and hydroacoustic small array. *Geophys. Res. Lett.* **32**, L15310, doi:10.1029/2005GL022966 (2005).
- de Groot-Hedlin, C. D. Estimation of the rupture length and velocity of the Great Sumatra earthquake of Dec 26, 2004 using hydroacoustic signals. *Geophys. Res. Lett.* **32**, L11303, doi:10.1029/2005GL022695 (2005).
- Lay, T. *et al.* The great Sumatra–Andaman earthquake of 26 December 2004. *Science* **308**, 1127–1133 (2005).
- Ammon, C. J. *et al.* Rupture process of the 2004 Sumatra–Andaman earthquake. *Science* **308**, 1133–1139 (2005).
- Prawirodirdjo, L. *et al.* Geodetic observations of interseismic strain segmentation at the Sumatra subduction zone. *Geophys. Res. Lett.* **24**, 2601–2604 (1997).
- McCaffrey, R. *et al.* Strain partitioning during oblique plate convergence in northern Sumatra: Geodetic and seismologic constraints and numerical modeling. *J. Geophys. Res.* **105**, 28363–28376 (2000).
- Vigny, C. *et al.* Insight into the 2004 Sumatra–Andaman earthquake from GPS measurements in southeast Asia. *Nature* **436**, 201–206 (2005).
- Nikolaidis, R. Observation of geodetic and seismic deformation with the Global Positioning System. PhD thesis, Univ. California, San Diego (2002).
- Altamimi, Z., Sillard, P. & Boucher, C. ITRF2000: A new release of the International Terrestrial Reference frame for earth science applications. *J. Geophys. Res.* **107** (B10), 2214, doi:10.1029/2001JB000561 (2002).
- Natawidjaja, D. H. *et al.* Paleogeodetic records of seismic and aseismic subduction from central Sumatran microatolls, Indonesia. *J. Geophys. Res.* **109**, doi:10.1029/2003JB0002398 (2004).
- Zachariassen, J., Sieh, K., Taylor, F. W. & Hantoro, W. S. Modern vertical deformation above the Sumatran subduction zone: Paleogeodetic insights from coral microatolls. *Bull. Seismol. Soc. Am.* **90**, 897–913 (2000).
- Agnew, D. C. NLOADF: a program for computing ocean-tide loading. *J. Geophys. Res.* **102**, 5109–5110 (1997).
- Meltzner, A. J. *et al.* Uplift and subsidence associated with the great Aceh–Andaman earthquake of 2004. *J. Geophys. Res.* **111**, doi:10.1029/2005JB003891 (2006).
- Jade, J., Ananda, M. B., Dileep Kumar, P. & Banerjee, S. Co-seismic and post-seismic displacements in Andaman and Nicobar Islands from GPS measurements. *Curr. Sci.* **88**, 1980–1984 (2005).
- Ruff, L. & Kanamori, H. Seismicity and the subduction process. *Phys. Earth Planet. Inter.* **23**, 240–252 (1980).
- Kanamori, H. in *Earthquakes: Observation, Theory and Interpretation* (eds Kanamori, H. & Bosch, E.) 597 (North Holland, New York, 1983).
- Uyeda, S. & Kanamori, H. Back-arc opening and the model of subduction. *J. Geophys. Res.* **84**, 1049–1061 (1979).
- Scholz, C. H. & Campos, J. On the mechanism of seismic decoupling and back-arc spreading at subduction zones. *J. Geophys. Res.* **100**, 22103–22115 (1995).
- McCaffrey, R. Influences of recurrence times and fault zone temperatures on the age-rate dependence of subduction zone seismicity. *J. Geophys. Res.* **102**, 22839–22854 (1997).
- Natawidjaja, D. H. *et al.* The giant Sumatran megathrust ruptures of 1797 and 1833: Paleoseismic evidence from coral microatolls. *J. Geophys. Res.* (in the press).
- Chlieh, M., Avouac, J., Sieh, K., Natawidjaja, D. & Galetzka, J. Investigating lateral variations of interseismic strain along the Sumatran subduction zone. *Eos* **85** (Fall Meeting Suppl.), abstr. T13C–1385 (2004).
- McCaffrey, R. Crustal block rotations and plate coupling. In *Plate Boundary Zones* (eds Stein, S. & Freymueller, J.) 101–122, doi:10.1029/030GD06 (AGU Geodynamics Series 30, Washington DC, 2002).
- Simoes, M., Avouac, J.-P., Cattin, R. & Henry, P. The Sumatra subduction zone: a case for a locked fault zone extending into the mantle. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002958 (2004).
- Bollinger, L., Avouac, J. P., Cattin, R. & Pandey, M. R. Stress buildup in the Himalaya. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002911 (2004).
- Miyazaki, S., Segall, P., Fukuda, J. & Kato, T. Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. *Geophys. Res. Lett.* **31**, doi:10.1029/2003GL019410 (2004).
- Cohen, S. C. On the rapid postseismic uplift along Turnagain Arm, Alaska following the 1964 Prince William Sound Earthquake. *Geophys. Res. Lett.* **25**, 1213–1215 (1998).
- Hutton, W., DeMets, C., Sanchez, O., Suarez, G. & Stock, J. Slip kinematics and dynamics during and after the 1995 October 9 $M_w = 8.0$ Colima–Jalisco earthquake, Mexico, from GPS geodetic constraints. *Geophys. J. Int.* **146**, 637–658 (2001).
- Hsu, Y. J. *et al.* Rapid afterslip following the 1999 Chi-Chi, Taiwan earthquake. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL014967 (2002).
- Scholz, C. *The Mechanics of Earthquakes and Faulting* Ch. 2 (Cambridge Univ. Press, New York, 1990).
- Hyndman, R. D., Yamano, M. & Oleskevich, D. A. The seismogenic zone of subduction thrust faults. *The Island Arc* **6**, 244–260 (1997).
- Pacheco, J. F., Sykes, L. R. & Scholz, C. H. Nature of seismic coupling along simple plate boundaries of the subduction type. *J. Geophys. Res.* **98**, 14133–14160 (1993).
- Brune, J. N. Seismic moment, seismicity, and rate of slip along major fault zones. *J. Geophys. Res.* **73**, 777–784 (1968).
- Curray, J. Tectonics and history of the Andaman Sea region. *J. Asian Earth Sci.* **25**, 187–232 (2005).
- Engdahl, E., van der Hilst, R. & Buland, R. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull. Seismol. Soc. Amer.* **88**, 722–743 (1998).
- Tsai, V. C., Nettles, M., Ekstrom, G. & Dziewonski, A. M. Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys. Res. Lett.* **32**, L17304, doi:10.1029/2005GL023813 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the team from BAKOSURTANAL who collected the GPS field data in Sumatra under difficult conditions (J. Efendi, A. Indrajit, M. Nyamadi, C. Bagandi, D. Sudharmono, A. Suryono, M. Achmad, U. Santoso, C. Yuniarsa, Endang, A. Pujubuntoro, H. Dradjat, B. Susilo and B. Parjanto). We are grateful to C. Vigny for comments and suggestions. This work has benefited from discussions with C. Ji. The Gordon and Betty Moore Foundation, the US National Science Foundation, the Southern California Earthquake Center, and BAKOSURTANAL supported this research. We thank R. W. Matindas and J. McRaney for their support. This is Caltech Tectonics Observatory contribution number 31.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-P.A. (avouac@gps.caltech.edu).

ARTICLES

Optimal isotope labelling for NMR protein structure determinations

Masatsune Kainosho¹, Takuya Torizawa¹, Yuki Iwashita¹, Tsutomu Terauchi¹, Akira Mei Ono¹ & Peter Güntert²

Nuclear-magnetic-resonance spectroscopy can determine the three-dimensional structure of proteins in solution. However, its potential has been limited by the difficulty of interpreting NMR spectra in the presence of broadened and overlapping resonance lines and low signal-to-noise ratios. Here we present stereo-array isotope labelling (SAIL), a technique that can overcome many of these problems by applying a complete stereospecific and regiospecific pattern of stable isotopes that is optimal with regard to the quality and information content of the resulting NMR spectra. SAIL uses exclusively chemically and enzymatically synthesized amino acids for cell-free protein expression. We demonstrate for the 17-kDa protein calmodulin and the 41-kDa maltodextrin-binding protein that SAIL offers sharpened lines, spectral simplification without loss of information, and the ability to rapidly collect the structural restraints required to solve a high-quality solution structure for proteins twice as large as commonly solved by NMR. It thus makes a large class of proteins newly accessible to detailed solution structure determination.

In the past two decades, nuclear-magnetic-resonance spectroscopy has become one of the two accepted methods (along with X-ray crystallography) for determining three-dimensional structures of proteins. NMR spectroscopy provides information about the structure and dynamic properties of proteins in solution, and offers an approach for determining the three-dimensional protein structures of systems that fail to crystallize. NMR currently provides about 15% of the protein structures in the Protein Data Bank and has a role in structural genomics¹, but complete automation of the structure determination process or the structural analysis of proteins with a molecular mass greater than 25 kDa have not yet become routine with NMR. Conventional uniform labelling of proteins with ¹³C and ¹⁵N, coupled with double and triple resonance, two-dimensional to four-dimensional NMR data collection, supports the determination of NMR solution structures of proteins, in favourable cases as large as 25 kDa (refs 2, 3). Cryogenic probes are used to improve the signal-to-noise ratio, and higher-field magnets provide increased resolution and a further gain in sensitivity. However, as molecular masses increase, NMR spectra become increasingly difficult to interpret because of spectral crowding and line broadening due to fast transverse relaxation. Spin-diffusion effects decrease the ability to determine inter-proton distances from nuclear Overhauser enhancement (NOE) data. A common approach for addressing these problems is to label proteins with deuterium to simplify the spectra and to minimize spin-diffusion effects^{4,5}. Data collection by transverse relaxation optimized spectroscopy (TROSY)⁶ provides sharper lines for amide and aromatic groups. With these methods the global folds of a limited number of proteins larger than 30 kDa could be derived on the basis of conformational restraints for amide and selected methyl groups⁷. However, conformational data from other aliphatic or aromatic groups and thus for most of the side chains remain difficult to collect for proteins larger than 25 kDa. Consequently, by June 2005 only 1% of all NMR protein structures deposited in the Protein Data Bank were for proteins with a molecular mass of more than 25 kDa, and backbone and side-chain chemical shift assignments more than 70% complete had been

recorded in the BioMagResBank (<http://www.bmrb.wisc.edu>) for only eight proteins larger than 25 kDa.

It has long been recognized that deuteration can be used to simplify NMR spectra⁸, to obviate the need for chiral assignments⁹, to facilitate the measurement of spin-spin⁹ and dipolar couplings⁷, and to increase resolution¹⁰. However, labelling patterns and the approaches for achieving them have so far been suboptimal. Random fractional deuteration methods⁹ suffer from the production of

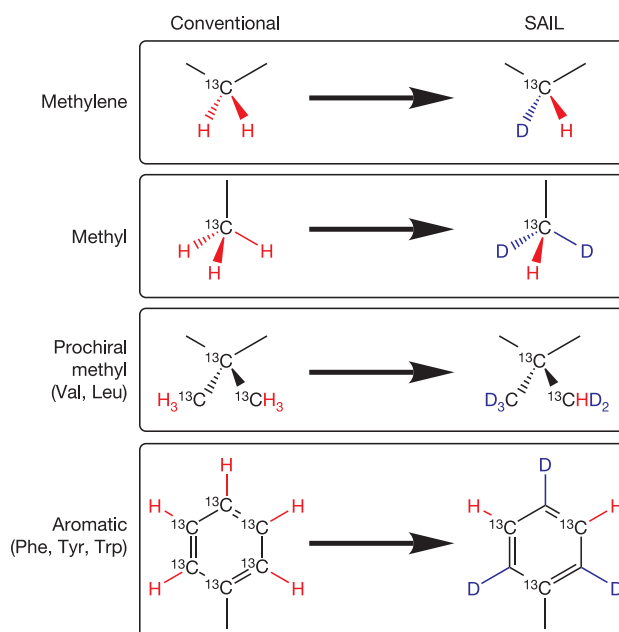


Figure 1 | SAIL amino acids. Design concepts embodied in the SAIL amino acids incorporated into CaM and MBP^{13–15}.

¹CREST/JST and Graduate School of Science, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji, 192-0397, Japan. ²Tatsuo Miyazawa Memorial Program, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, 230-0045, Japan.

Table 1 | Expected and observed features of CaM and MBP samples prepared by conventional uniform labelling and SAIL methods

Quantity	CaM		MBP	
	Uniform labelling	SAIL	Uniform labelling	SAIL
¹ H atoms per molecule in ¹ H ₂ O	1,095	697 (64%)	2,860	1,802 (63%)
¹ H atoms per molecule in ² H ₂ O*	851	453 (53%)	2,249	1,191 (53%)
Non-exchangeable side-chain ¹ H atoms	692	305 (44%)	1,850	821 (44%)
NOE cross-peaks expected†	9,812	5,642 (58%)	17,076	9,382 (55%)
NOE cross-peaks used	–	4,576 (47%)	–	7,485 (44%)
NOE distance restraints expected‡	2,883	2,720 (94%)	4,293	4,347 (101%)
NOE distance restraints used	–	2,422 (84%)	–	3,818 (89%)

*The number was calculated by assuming that all exchangeable ¹H nuclei are replaced by ²H.

†NOE cross-peaks are assumed to be observable between all aliphatic, aromatic, backbone amide and Asn/Gln side-chain amide ¹H nuclei that are less than 4.75 Å apart in the 20 conformers of the SAIL-CaM solution structure, or less than 4.2 Å apart in the 20 conformers of the SAIL-MBP solution structure.

‡Excluding duplicate restraints from symmetry-related peak pairs and NOEs for fixed distances. For uniform labelling, pseudo-atoms were assumed.

numerous isotopomer proteins with chemical shift heterogeneity and decreased signal intensities. Perdeuteration methods remove all carbon-bound protons such that much potential NOE information is lost. Methods for introducing methyl and/or aromatic protons into a perdeuterated background improve this situation, but the additional structural information is localized and unevenly distributed^{4,5}. The same holds true for residue-selective¹¹ and segmental¹² labelling. As a consequence, the quality of three-dimensional protein structures based on these methods remains limited. As an alternative presented here, optimal labelling patterns for protein NMR can be realized by the chemical or enzymatic synthesis of amino acids^{13–15} followed by *in vitro* (cell-free) protein expression¹⁶ to build a protein exclusively from such amino acids. Cell-free methods^{17,18} are crucial to making efficient use of the labelled amino acids and to the prevention of scrambling of the label, which would occur through metabolic pathways present in cell-based protein expression systems^{11,19}.

Stereo-array isotope labelling (SAIL)

The basic strategy of the SAIL approach is to prepare amino acids with the following features (Fig. 1): first, stereo-selective replacement of one ¹H in methylene groups by ²H; second, replacement of two ¹H in each methyl group by ²H; third, stereo-selective modification of the prochiral methyl groups of Leu and Val such that one methyl is –¹²C(²H)₃ and the other is –¹³C¹H(²H)₂; and last, labelling of six-membered aromatic rings by alternating ¹²C–²H and ¹³C–¹H moieties.

This labelling pattern preserves through-bond connectivity information needed for backbone and side chain assignments, eliminates the need for stereospecific assignments, simplifies measurements of couplings, and removes the most serious sources of spin diffusion so as to improve the accuracy of inter-proton distance measurements. Lines are sharpened both by decreasing long-range couplings and by eliminating dipolar relaxation pathways. The methyl and methylene labelling patterns simplify the analysis of the motional properties of the side chains from relaxation measurements²⁰. The aromatic ring labelling strategy removes one-bond ¹³C–¹³C couplings, which often complicate spectra or require the use of constant time data collection methods so as to reduce spectral complexity²¹.

We prepared the 20 protein-component SAIL amino acids (Supplementary Fig. 1) based on these design concepts by chemical and enzymatic syntheses^{13–15}. Efficient incorporation of these amino acids into the protein of interest is achieved by an *Escherichia coli* cell-free protein synthesis system optimized for the preparation of labelled NMR samples¹⁶. We chose as our first targets for SAIL the 17-kDa calcium-binding protein calmodulin (CaM) from *Xenopus laevis* and the 41-kDa maltodextrin-binding protein (MBP) from *E. coli*. SAIL amino acid mixture (55 mg), prepared by following the protocol¹⁶, yielded 5.5 mg of purified, soluble SAIL-CaM, or 5.3 mg of purified, soluble SAIL-MBP. For comparison, CaM and MBP uniformly labelled with ¹³C and ¹⁵N (UL-CaM and UL-MBP, respectively) were also prepared.

SAIL versus uniform labelling with ¹³C and ¹⁵N

On the basis of calculations (Table 1), we expected that the NMR spectra of SAIL proteins would be simpler than those of the corresponding uniformly labelled proteins. The numbers of observable protons are reduced by SAIL relative to uniform labelling. The SAIL approach reduces the number of non-exchangeable side-chain protons, which are prone to overlap but essential for defining the side-chain conformations, to less than half, and decreases the number of expected NOESY cross-peaks by 40–45%. Most of the additional NOEs from uniformly labelled proteins either involve fixed (geminal) distances or become redundant in the absence of stereospecific assignments and thus contribute to spectral overlap without furnishing independent information. In principle, stereospecific assignments could allow slightly more precise structures in the case of uniform labelling^{22,23}. However, in practice only a limited number of stereospecific assignments can be made^{22–24}. The BioMagResBank reports stereospecific assignments for 7% of the methylene protons and Val/Leu methyl groups, and for only 3.4% of the prochiral groups in proteins larger than 20 kDa. Therefore, in general the expected number of non-redundant, structurally relevant NOE restraints is retained with SAIL (Table 1), even if the effect of the better signal-to-noise ratio with SAIL is neglected. These numbers are corroborated by the experimental findings (Supplementary Table 1). More detailed theoretical considerations (see Methods), which take into account the enhanced signal strength and sharper lines in SAIL spectra and the fact that overlap can render peaks unidentifiable, show that in practice SAIL is expected to increase rather than decrease the number of identifiable NOE cross-peaks. The expected increase is moderate in regions without overlap but is significant in regions with strong overlap and therefore for larger proteins, for which SAIL is expected to yield two or more times the number of relevant conformational restraints than uniform labelling.

Precautions to be taken when collecting and analysing data for a SAIL protein are that deuterium decoupling should be applied during ¹³C evolution times and that what normally are methyl and methylene groups should be treated as methine groups. ¹H–¹³C CT-HSQC (constant-time heteronuclear single-quantum coherence) NMR data sets collected from SAIL and uniformly labelled proteins demonstrate the superiority of the SAIL method (Figs 2 and 3). Each ¹H–¹³C pair in the protein is associated with a peak in the spectrum. Severe signal overlap in uniform labelling is alleviated in the spectrum from the SAIL proteins. The improvements, which are particularly apparent for larger proteins (Fig. 3), result from the decrease in the number of ¹H signals as well as from sharpening of the remaining signals.

The SAIL method also improves sensitivity. Part of the gain arises from longer ¹H and ¹³C transverse relaxation times resulting from replacements of ¹H by ²H. Reduced relaxation during magnetization transfer steps in experiments such as ¹H–¹³C CT-HSQC leads to an increased signal-to-noise ratio. Reduced long-range couplings result in a further sharpening of signals. The signal intensities for methylene

groups are threefold to sevenfold higher with SAIL than with uniform labelling under the same conditions. Improvements are more pronounced for the 41-kDa SAIL-MBP protein (Fig. 3e) than for the 17-kDa SAIL-CaM (Fig. 2i). Although each observed SAIL methyl group contained only one ^1H in comparison with three equivalent ^1H with uniform labelling, equivalent signal intensities were observed (data not shown) as a result of the longer ^1H and ^{13}C transverse relaxation times.

SAIL protein structure determinations

All peaks benefited from improved sensitivity and resolution and allowed the signals of SAIL-CaM and SAIL-MBP to be assigned readily by established methods²⁵. Side-chain assignments were determined completely from the analysis of two data sets: HCCH-TOCSY (total correlation spectroscopy) data provided connectivities among all side-chain signals, and HCCH-COSY (correlation spectroscopy) data were used to identify the spin systems. The detection and assignment of signals from aromatic rings containing alternating ^{12}C and ^{13}C nuclei (Fig. 1) is straightforward by an unconventional approach that we describe separately²⁶. The expected ^1H , ^{13}C and ^{15}N chemical shifts could be assigned without exception for SAIL-CaM, and to 94% for SAIL-MBP, including more than 90% of the aliphatic and aromatic side-chain protons (Supplementary Table 1). Many of the shifts that could not be assigned are in the region of residues

229–241, which have been shown to interact with the bound cyclodextrin²⁷. Because of conformational heterogeneity, most of the expected resonances of residues 229–241 could also not be assigned in earlier studies⁷.

We obtained distance restraints for the structure calculation from three-dimensional ^{15}N - and ^{13}C -edited NOESY spectra²⁸ and from a two-dimensional NOESY spectrum for the aromatic region. These spectra were simplified by a decreased number of signals, as expected (Table 1). Because of lower spin diffusion, maximum NOE intensities for SAIL proteins were typically reached at mixing times 1.5-fold to 3.0-fold those for uniform labelling (data not shown). On the basis of the simplified and more quantitative NOESY spectra, we were able to obtain structures of SAIL-CaM and SAIL-MBP by means of the combined automated NOE assignment and structure calculation protocol in the program CYANA^{29,30} (Supplementary Table 1). For instance, a dense network of NOEs including 949 non-redundant, long-range distance restraints was established for SAIL-MBP. Of the 3,818 non-redundant NOE distance restraints, 1,879 involve side-chain atoms beyond H^β . The SAIL-CaM and SAIL-MBP solution structures show good quality in terms of the agreement with the experimental data and other validation parameters.

Structures of the calcium-bound form of calmodulin have been determined previously by crystallography³¹ and NMR. The NMR structure³² was based on residual dipolar couplings (RDCs),

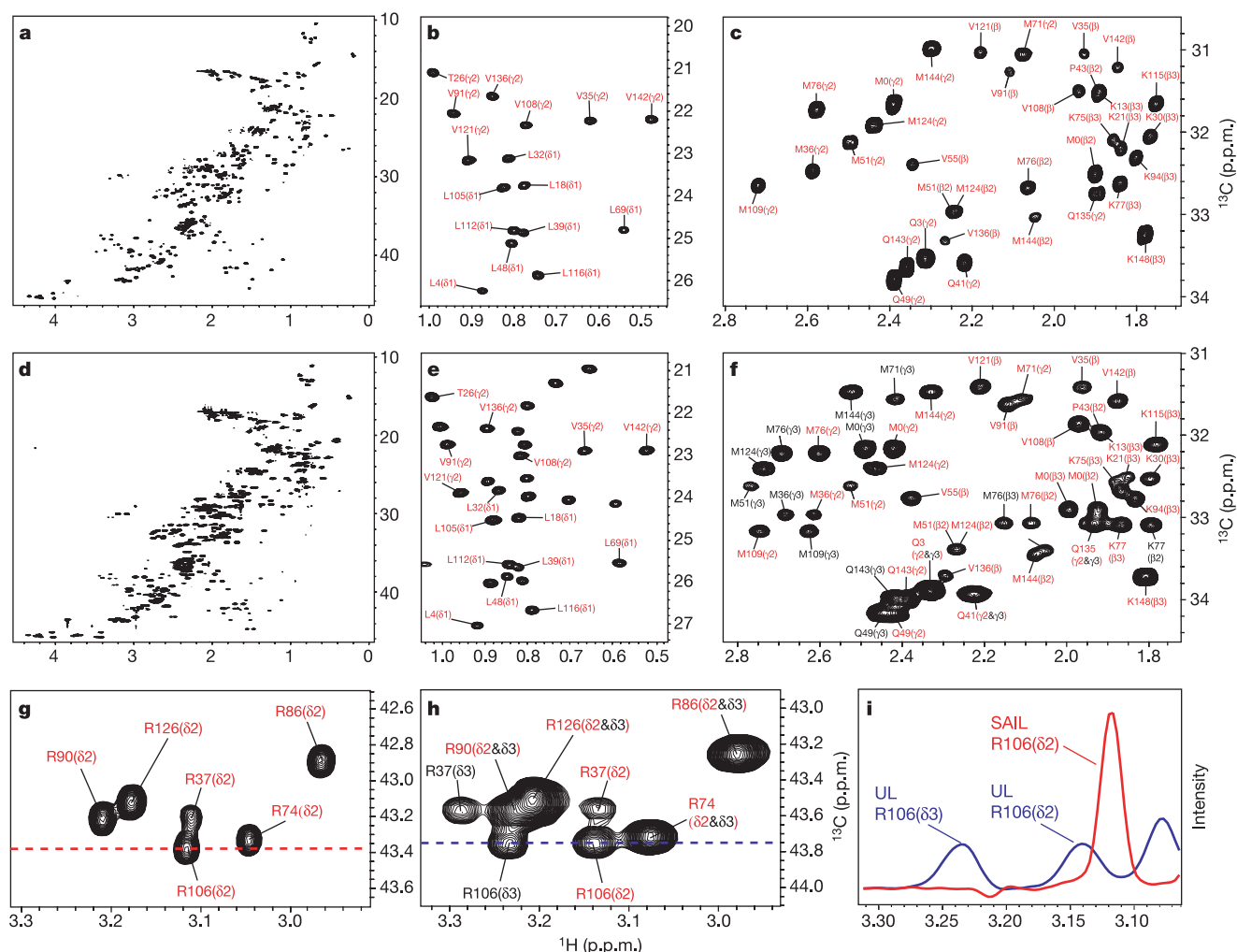


Figure 2 | ^1H - ^{13}C CT-HSQC spectra of CaM. **a**, SAIL-CaM, aliphatic region. **b**, SAIL-CaM, methyl region. **c**, SAIL-CaM, methylene region. **d**, UL-CaM, aliphatic region. **e**, UL-CaM, methyl region. **f**, UL-CaM, methylene region. **g**, SAIL-CaM, Arg δ region. **h**, UL-CaM, Arg δ region. **i**, Cross-sections from

g and **h**. The spectra for SAIL-CaM and UL-CaM were recorded under identical conditions and scaled for equal noise levels. Assignments are indicated by one-letter amino-acid code, residue number and atom identifier. Assignments of UL-CaM are as reported previously³⁷.

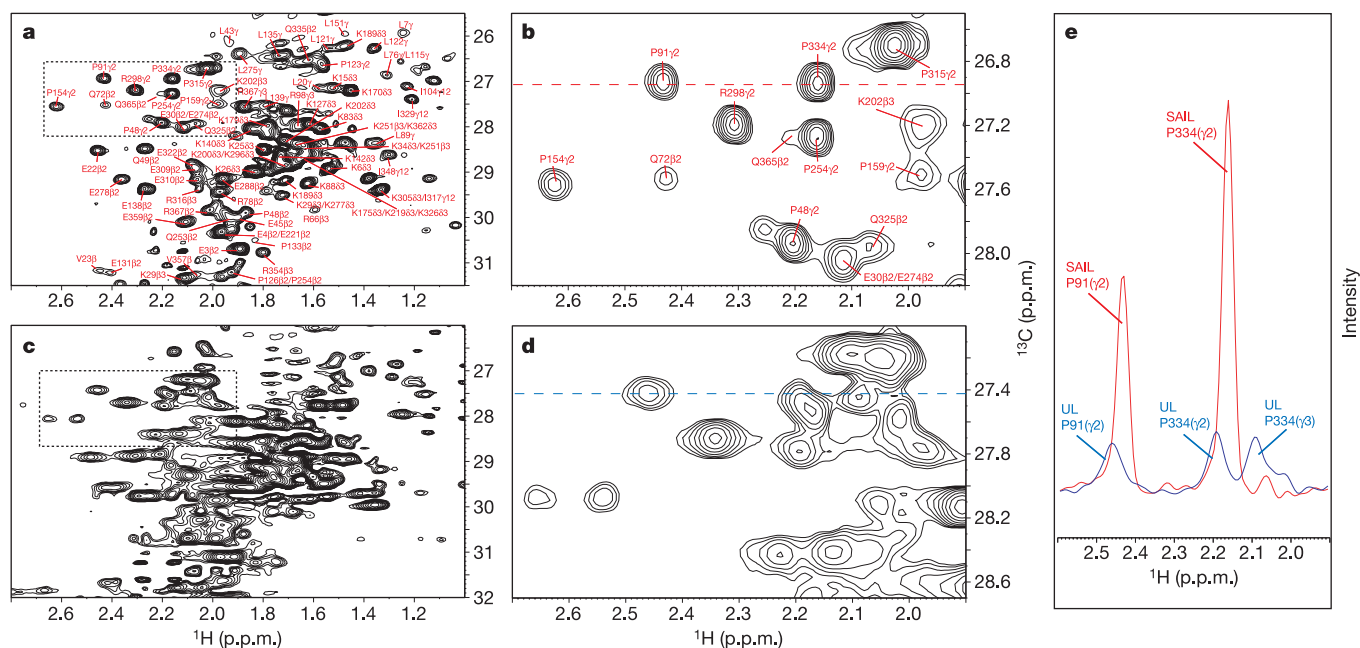


Figure 3 | ^1H - ^{13}C CT-HSQC spectra of MBP. **a**, Aliphatic region of CDH groups in SAIL-MBP. **b**, Enlargement of the rectangular region marked in **a**. Assignments are indicated with one-letter amino-acid code, residue number and atom identifiers. **c**, **d**, Corresponding regions for UL-MBP.

e, Cross-sections taken at the positions indicated in **b** and **d**. The spectra for SAIL-MBP and UL-MBP were recorded under identical conditions and scaled for equal noise levels.

measured mainly for the polypeptide backbone, and on the crystal structure. Our NMR structure, the first one for calcium-bound calmodulin based exclusively on NOEs, provides detailed information on the side-chain conformations. It is in close agreement with the crystal structure and the RDC-based backbone structure (Fig. 4a).

In addition, the solution structure of the 41-kDa SAIL-MBP coincides closely with the crystal structure²⁷ determined previously under slightly different conditions (Fig. 4b, c, and Supplementary Table 1). The 41-kDa SAIL-MBP solution structure is of similar precision and accuracy to those of smaller proteins, and the structural statistics are comparable to those commonly found in NMR

structure determinations of smaller proteins. Previously⁷, a global fold of MBP was determined by NMR on the basis of NOEs between amide and methyl protons, residual dipolar couplings for the polypeptide backbone, and hydrogen-bond restraints. That NMR study could provide a good determination of the global fold of the polypeptide backbone and the conformations of the methyl-containing side chains of valine, leucine and isoleucine, but the approach used cannot provide direct structural information on the other side chains. Their conformations therefore remained largely undetermined, resulting in root-mean-square deviations (RMSDs) of more than 3.8 Å for all side-chain heavy atoms of the amino-terminal and carboxy-terminal domains of MBP. The corresponding

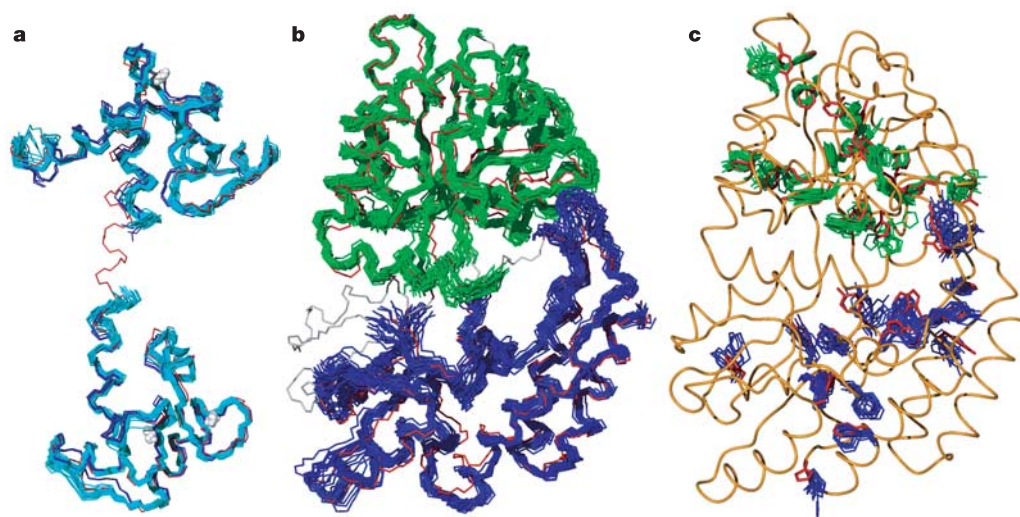


Figure 4 | CaM and MBP solution and crystal structures. **a**, SAIL-CaM (backbone in cyan, Ca^{2+} in white), CaM X-ray structure³¹ (red), and three solution conformers of UL-CaM determined from residual dipolar coupling data³² (blue). **b**, MBP solution and crystal structures²⁷. Backbone of the N-terminal domain (SAIL-MBP in green, X-ray in red) and the

C-terminal domain (SAIL-MBP in blue, X-ray in red). **c**, Aromatic side chains (SAIL-MBP in green and blue, X-ray in red), and backbone ribbon representation of the X-ray structure (gold). Superpositions of CaM and MBP solution conformers on the X-ray structures were performed separately for the two flexibly connected domains.

side-chain RMSDs for the SAIL-MBP structure are 2.3 Å for both domains.

Discussion

The SAIL strategy described here is expected to support high-throughput protein structure determination without loss of structural quality. The much smaller numbers of ^1H shifts that need to be assigned make SAIL proteins particularly amenable to automated resonance assignment, which also benefits from the complete absence of uncertainties associated with the lack of stereospecific assignments and frequent accidental chemical shift degeneracies for diastereotopic pairs in uniformly labelled proteins. SAIL improves the quality of virtually all commonly used multi-dimensional NMR spectra and can be used in conjunction with other techniques to improve the sensitivity of NMR experiments, such as TROSY⁶, cryogenic probes and high-field magnets. Our structure determination of MBP shows that high-quality solution structures of proteins up to at least 40 kDa can now be solved by NMR.

METHODS

Expected numbers of identifiable NOESY peaks with SAIL and uniform labelling.

We define a peak as identifiable if it is above the noise level and not overlapped with other peaks. Assuming N peaks, distributed randomly within a region of size Γ in a d -dimensional spectrum, the expected number of peaks that are not overlapped with other peaks is $N_1 = N(1 - \gamma/\Gamma)^{N-1} \approx N e^{-N\gamma/\Gamma}$, where γ denotes the size of the peak region for each peak. SAIL reduces the number of ^1H nuclei relative to uniform labelling, and thus yields a smaller number, $\tilde{N}_0 < N$, of NOE signals with integral above the same threshold. Reduced relaxation gives rise to narrower line widths, $\Delta\tilde{\omega} < \Delta\omega$, greater peak heights, $\tilde{H} = H(\Delta\omega/\Delta\tilde{\omega})^d$, smaller peak areas, $\tilde{\gamma} = \gamma(\Delta\tilde{\omega}/\Delta\omega)^d$, and an increased number of peaks above the noise level, $\tilde{N} = \tilde{N}_0(\Delta\omega/\Delta\tilde{\omega})^{d/2}$. The latter follows from the relationship $N(H \geq H_{\min}) = N(I \geq I_{\min}) = N(r \leq r_{\max}) \propto r_{\max}^3 \propto I_{\min}^{-1/2} \propto \Delta\omega^{-d/2}$, where H , I and r are respectively the peak height, peak integral and distance for a given NOE, and $H_{\min}$, $I_{\min}$ and $r_{\max}$ are the noise level, corresponding intensity threshold and maximal NOE observation distance, respectively. The number of identifiable peaks for a SAIL protein, $\tilde{N}_1$, becomes, in terms of the corresponding number, N_1 , for the uniformly labelled protein,

$$\tilde{N}_1 \approx \tilde{N} e^{-\tilde{N}\tilde{\gamma}/\Gamma} = N_1 \sqrt{\frac{\tilde{H}}{H}} \left(\frac{\tilde{N}_0}{N} \right) \exp \left[\left(1 - \sqrt{\frac{\tilde{H}}{H}} \left(\frac{\tilde{N}_0}{N} \right) \frac{N\gamma}{\Gamma} \right) \right] \left(\frac{\tilde{N}_0}{N} \right)$$

For MBP with a SAIL-to-UL peak height ratio of $\tilde{H}/H \approx 4$ –7 (Fig. 3e) and a ratio $\tilde{N}_0/N \approx 0.55$ (Table 1) for the number of NOEs below 4.2 Å, SAIL is expected to increase the number of identifiable NOEs by at least 10% for regions without overlap ($N\gamma/\Gamma = 0$) and by more than 50% for a crowded region with $N\gamma/\Gamma = 0.5$.

NMR spectroscopy. The SAIL-CaM and UL-CaM samples each contained 0.7 mM protein, 5 mM MES-d13 and 10 mM bis-Tris-d19 (Cambridge Isotope Laboratories), 5 mM CaCl_2 and 0.1 mM NaN_3 , pH 6.5. The SAIL-MBP and UL-MBP samples each contained 0.33 mM protein, 3.3 mM β -cyclodextrin, 20 mM sodium phosphate, 3 mM NaN_3 and Complete Mini protease inhibitor mix (Roche), pH 7.2. NMR experiments were performed at 37 °C on a Bruker DRX800 spectrometer equipped with a TXI xyz-gradient probe. NMR spectra were analysed with the program SPARKY (University of San Francisco, California).

Structure calculation. The SAIL-CaM and SAIL-MBP structures were obtained with the program CYANA²⁹ with the use of automated NOE assignment³⁰ and torsion angle dynamics for the structure calculation³³, which was started from 100 (CaM) or 200 (MBP) conformers with random torsion angle values. The standard CYANA-simulated annealing schedule was applied with 10,000 (CaM) or 20,000 (MBP) torsion angle dynamics steps²⁹. Backbone torsion angle restraints obtained from chemical shifts with the program TALOS³⁴ were added to the input for CYANA. Hydrogen-bond restraints were not used. The 20 conformers with the lowest final CYANA target function values were embedded in a water shell of 8 Å thickness and energy-minimized against the AMBER force field³⁵ with the program OPALp³⁶ in the presence of the NOE distance restraints as the only experimental data.

Received 8 September; accepted 14 December 2005.

- Kennedy, M. A., Montelione, G. T., Arrowsmith, C. H. & Markley, J. L. Role for NMR in structural genomics. *J. Struct. Funct. Genom.* **2**, 155–169 (2002).
- Clore, G. M. & Gronenborn, A. M. Structures of larger proteins, protein–ligand

and protein–DNA complexes by multidimensional heteronuclear NMR. *Prog. Biophys. Mol. Biol.* **62**, 153–184 (1994).

- Clore, G. M. & Gronenborn, A. M. NMR structure determination of proteins and protein complexes larger than 20 kDa. *Curr. Opin. Chem. Biol.* **2**, 564–570 (1998).
- Gardner, K. H. & Kay, L. E. The use of ^2H , ^{13}C , ^{15}N multidimensional NMR to study the structure and dynamics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 357–406 (1998).
- Goto, N. K. & Kay, L. E. New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Curr. Opin. Struct. Biol.* **10**, 585–592 (2000).
- Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T_2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA* **94**, 12366–12371 (1997).
- Mueller, G. A. *et al.* Global folds of proteins with low densities of NOEs using residual dipolar couplings: Application to the 370-residue maltodextrin-binding protein. *J. Mol. Biol.* **300**, 197–212 (2000).
- Markley, J. L., Putter, I. & Jardetzky, O. High-resolution nuclear magnetic resonance spectra of selectively deuterated staphylococcal nuclease. *Science* **161**, 1249–1251 (1968).
- LeMaster, D. M. Chiral β and random fractional deuteration for the determination of protein side-chain conformation by NMR. *FEBS Lett.* **223**, 191–196 (1987).
- Farmer, B. T. II & Venter, R. A. NMR of perdeuterated large proteins. *Mod. Techniques Protein NMR* **16**, 75–120 (1999).
- Arata, Y., Kato, K., Takahashi, H. & Shimada, I. Nuclear-magnetic-resonance study of antibodies—a multinuclear approach. *Methods Enzymol.* **239**, 440–464 (1994).
- Yamazaki, T. *et al.* Segmental isotope labeling for protein NMR using peptide splicing. *J. Am. Chem. Soc.* **120**, 5591–5592 (1998).
- Oba, M., Kobayashi, M., Oikawa, F., Nishiyama, K. & Kainosho, M. Synthesis of $^{13}\text{C}/\text{D}$ doubly labelled L-leucines: Probes for conformational analysis of the leucine side-chain. *J. Org. Chem.* **66**, 5919–5922 (2001).
- Oba, M., Terauchi, T., Miyakawa, A., Kamo, H. & Nishiyama, K. Stereoselective deuterium-labelling of diastereotopic methyl and methylene protons of L-leucine. *Tetrahedron Lett.* **39**, 1595–1598 (1998).
- Oba, M., Terauchi, T., Miyakawa, A. & Nishiyama, K. Asymmetric synthesis of L-proline regio- and stereoselectively labelled with deuterium. *Tetrahedron Asymmetry* **10**, 937–945 (1999).
- Torizawa, T., Shimizu, M., Taoka, M., Miyano, H. & Kainosho, M. Efficient production of isotopically labelled proteins by cell-free synthesis: A practical protocol. *J. Biomol. NMR* **30**, 311–325 (2004).
- Kigawa, T., Muto, Y. & Yokoyama, S. Cell-free synthesis and amino acid-selective stable-isotope labeling of proteins for NMR analysis. *J. Biomol. NMR* **6**, 129–134 (1995).
- Zubay, G. In-vitro synthesis of protein in microbial systems. *Annu. Rev. Genet.* **7**, 267–287 (1973).
- McIntosh, L. P. & Dahlquist, F. W. Biosynthetic incorporation of ^{15}N and ^{13}C for assignment and interpretation of nuclear-magnetic-resonance spectra of proteins. *Q. Rev. Biophys.* **23**, 1–38 (1990).
- Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. A. Pulse sequences for removal of the effects of cross-correlation between dipolar and chemical-shift anisotropy relaxation mechanism on the measurement of heteronuclear T_1 and T_2 values in proteins. *J. Magn. Reson.* **97**, 359–375 (1992).
- Vuister, G. W. & Bax, A. Resolution enhancement and spectral editing of uniformly ^{13}C -enriched proteins by homonuclear broadband ^{13}C decoupling. *J. Magn. Reson.* **98**, 428–435 (1992).
- Güntert, P., Braun, W., Billeter, M. & Wüthrich, K. Automated stereospecific ^1H NMR assignments and their impact on the precision of protein structure determinations in solution. *J. Am. Chem. Soc.* **111**, 3997–4004 (1989).
- Nilges, M., Clore, G. M. & Gronenborn, A. M. ^1H -NMR stereospecific assignments by conformational data-base searches. *Biopolymers* **29**, 813–822 (1990).
- Neri, D., Szyperski, T., Otting, G., Senn, H. & Wüthrich, K. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28**, 7510–7516 (1989).
- Cavanagh, J., Palmer, A. G., Fairbrother, W. & Skelton, N. *Protein NMR Spectroscopy: Principles and Practice* (Academic, San Diego, 1996).
- Torizawa, T., Ono, M. A., Terauchi, T. & Kainosho, M. NMR assignment methods for the aromatic ring resonances of phenylalanine and tyrosine residues in proteins. *J. Am. Chem. Soc.* **127**, 12620–12626 (2005).
- Sharff, A. J., Rodseth, L. E. & Quiocho, F. A. Refined 1.8-Å structure reveals the mode of binding of β -cyclodextrin to the maltodextrin binding protein. *Biochemistry* **8**, 10553–10559 (1993).
- Marion, D., Kay, L. E., Sparks, S. W., Torchia, D. A. & Bax, A. 3-dimensional heteronuclear NMR of ^{15}N -labelled proteins. *J. Am. Chem. Soc.* **111**, 1515–1517 (1989).
- Güntert, P. Automated NMR protein structure calculation. *Prog. NMR Spectrosc.* **43**, 105–125 (2003).

30. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227 (2002).
31. Chattopadhyaya, R., Meador, W. E., Means, A. R. & Quirocho, F. A. Calmodulin structure refined at 1.7 Å resolution. *J. Mol. Biol.* **228**, 1177–1192 (1992).
32. Chou, J. J., Li, S., Klee, C. B. & Bax, A. Solution structure of Ca²⁺-calmodulin reveals flexible hand-like properties of its domains. *Nature Struct. Biol.* **8**, 990–997 (2001).
33. Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997).
34. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302 (1999).
35. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
36. Koradi, R., Billeter, M. & Güntert, P. Point-centered domain decomposition for parallel molecular dynamics simulation. *Comput. Phys. Commun.* **124**, 139–147 (2000).
37. Ikura, M. *et al.* Secondary structure and side-chain ¹H and ¹³C resonance assignments of calmodulin in solution by heteronuclear multidimensional NMR spectroscopy. *Biochemistry* **30**, 9216–9228 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Ikura and T. Yamazaki for providing the calmodulin and maltodextrin-binding protein genes, respectively, and M. Takeda for help with the preparation of figures. This work was supported by CREST/JST.

Author Information Atomic coordinates of the SAIL-CaM and SAIL-MBP structures have been deposited in the Protein Data Bank with accession codes 1X02 and 2D21, respectively. Chemical shifts have been deposited in the BioMagResBank with accession numbers 6541 and 6807. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.K. (kainosho@nmr.chem.metro-u.ac.jp).

LETTERS

A magnetically collimated jet from an evolved star

Wouter H. T. Vlemmings¹, Philip J. Diamond¹ & Hiroshi Imai²

Planetary nebulae often have asymmetric shapes, even though their progenitor stars were symmetric; this structure could be the result of collimated jets from the evolved stars before they enter the planetary nebula phase^{1–3}. Theoretical models have shown that magnetic fields could be the dominant source of jet-collimation in evolved stars^{4,5}, just as these fields are thought to collimate outflows in other astrophysical sources, such as active galactic nuclei^{6–9} and proto-stars^{10,11}. But hitherto there have been no direct observations of both the magnetic field direction and strength in any collimated jet. Here we report measurements of the polarization of water vapour masers that trace the precessing jet emanating from the asymptotic giant branch star W43A (at a distance of 2.6 kpc from the Sun), which is undergoing rapid evolution into a planetary nebula^{2,12}. The masers occur in two clusters at opposing tips of the jets, $\sim 1,000$ AU from the star. We conclude from the data that the magnetic field is indeed collimating the jet.

The 22 GHz H₂O masers can be observed at high angular resolution and can be used to trace the magnetic field strength and structure at small scales^{13–15}. Circular polarization is caused by the Zeeman effect, observations of which yield the magnetic field strength along the line of sight^{13,14,16}. The linear polarization vectors describe the direction of the magnetic field.

We have used the Very Long Baseline Array (VLBA) of the National Radio Astronomy Observatory to determine both the linear and circular polarization of the 22 GHz H₂O masers in the jet of W43A (Figs 1 and 2). In our linear polarization spectra we achieve a r.m.s. noise of approximately 10 mJy beam^{-1} , and we detect linear polarization in the six brightest maser features (peak flux $> 6 \text{ Jy beam}^{-1}$), with a weighted mean linear polarization fraction of $0.66 \pm 0.07\%$. The 3σ r.m.s. upper limits of the linear polarization fraction on the other 14 detected but weaker features range upwards from 0.68%. The linear polarization vectors are predominantly aligned with the jet. According to maser theory, the linear polarization vectors are either parallel or perpendicular to the magnetic field direction¹⁷. The strongest linear polarization is typically found when the polarization vector is perpendicular to the magnetic field direction on the sky^{13,16}. We thus conclude that most of the observed linear polarization vectors are perpendicular to the magnetic field direction, as shown in Fig. 1c and d.

Across the brightest maser feature we observe a 90° flip of the polarization angle. Here, as predicted by the theory and as previously observed in SiO masers¹⁸, the linear polarization vector changes from parallel to perpendicular with respect to the magnetic field across the maser. We find the median, error weighted, position angle of the magnetic field to be $-27 \pm 12^\circ$ east of north. This error estimate includes the systematic error (8°) introduced after polarization calibration with respect to the calibrator J1743 – 0350. The magnetic field direction is thus almost perfectly perpendicular to the jet, which is consistent with a toroidal magnetic field configuration. We have also detected a fractional circular polarization

$P_V = 0.33 \pm 0.09\%$ in one of the maser features in the southern tip of the collimated jet (Fig. 2). Using a full 22 GHz H₂O maser radiative transfer code, which does not assume thermal equilibrium, to fit to the circular polarization^{13–16}, we find the magnetic field strength along the line of sight in the maser $B_{\parallel} = B \cos \theta = 85 \pm 33$ mG, where θ is the angle between the line of sight and the magnetic field direction. We estimate $\theta \approx 65^\circ$ using the linear polarization fraction ($0.64 \pm 0.19\%$) of the maser and our polarization model fit results. Thus, the magnetic field in the H₂O masers at one position within the collimated jet is $B \approx 200 \pm 75$ mG. This implies that, depending on the hydrogen number density in the H₂O maser regions of the jet, the magnetic pressure dominates the gas pressure by a factor of 2–200, as expected in the case of magnetic collimation⁵.

H₂O masers at 22 GHz are excited in gas with a hydrogen number density $n \approx 10^8$ – 10^{10} cm^{-3} (ref. 19). The H₂O masers in the collimated jet of W43A probably arise when the jet has swept up enough material previously expelled from the star so that conditions at the tip of the jet have become favourable for H₂O masers to occur². As the typical hydrogen number density at 1,000 AU from the star is $n \approx 10^5 \text{ cm}^{-3}$, the swept up material has increased the density in the maser region by a factor of between $\sim 10^3$ and 10^5 . When the magnetic field is partially coupled to the gas, compression of the magnetic field lines increases the field strength in the maser region, with the relation between the magnetic field strength (B) and density being $B \propto n^k$. Theoretical and observational considerations indicate that the coefficient k can range between 0.3 and 1. An empirical relation, which has $k = 0.47$, was determined from observations of magnetic fields in star-forming regions²⁰. Using this relation, the magnetic field strength in the lower density region outside the jet is $B \approx 0.9$ – 2.6 mG. This is consistent with values measured in the OH maser region at similar distances from the star in other systems. Alternatively, the masers may occur in a shock, similar to the H₂O masers found in star-forming regions²¹. Such shocks might exist between the collimated jet and dense material in the outer circumstellar envelope. For a shock model we calculate, using H₂O shock excitation models²¹, that the pre-shock hydrogen density is $\sim 3 \times 10^6 \text{ cm}^{-3}$ and the pre-shock magnetic field strength is $B \approx 0.07$ mG.

A rotating magnetic field in the outflow of a star can be described by two components: a toroidal component B_ϕ ($\propto r^{-1}$) and the radial component of the poloidal field B_r ($\propto r^{-2}$). Owing to the dependences on distance (r) from the star, the radial component can be effectively neglected at distances of 1,000 AU. This is consistent with the observed magnetic field directions, which are fully toroidal. Using the r^{-1} dependence, we can estimate the magnetic field strength B_ϕ at the base of the collimated jet on the surface of the star (with stellar radius $R \approx 1$ AU) to be ~ 1.5 G if the jet H₂O masers are excited in swept up material, or ~ 70 mG if the H₂O masers occur in shocks. The stellar toroidal magnetic field strength has a latitude dependence of the form $B_\phi(\theta_s) = B_s \sin \theta_s$, where B_s is the strength at the equator, $\theta_s = \pi/2$. Assuming we measure the magnetic field at the

¹Jodrell Bank Observatory, University of Manchester, Macclesfield, Cheshire SK11 9DL, UK. ²Department of Physics, Faculty of Science, Kagoshima University, Kagoshima 890-0065, Japan.

edge of the collimated jet with an opening angle of $\sim 5^\circ$, this implies $B_s \approx 35$ or 1.6 G. The higher value is in excellent agreement with measurements extrapolated from previous maser polarization observations^{14,15,18,22}.

The magnetic field and jet characteristics of W43A support the recent theoretical models that use magnetically collimated jets from evolved stars to explain the shaping of asymmetric planetary nebulae^{5,23,24}. These models also explain the shapes of a large number of other proto-planetary nebulae, such as He 3-401 (ref. 25). However, the origin of the stellar magnetic field is still a matter of debate. Magnetic dynamo models invoking the differential rotation

between a rapidly rotating core and a more slowly rotating outer layer have been shown to be able to produce magnetic fields in evolved stars similar in strength to those found in our observations⁴. Other models stress the need for a binary companion or heavy planet for a sufficient magnetic field to be attained²⁶. The presence of a heavy planet in orbit around W43A could also be the cause of the observed jet precession^{2,27}.

Different models have described the collimation of jets by magnetic fields around active galactic nuclei (AGN)⁷ and proto-stars²⁸. These models have been shown to be able to produce many of the observed characteristics of sources, such as the collimated proto-star

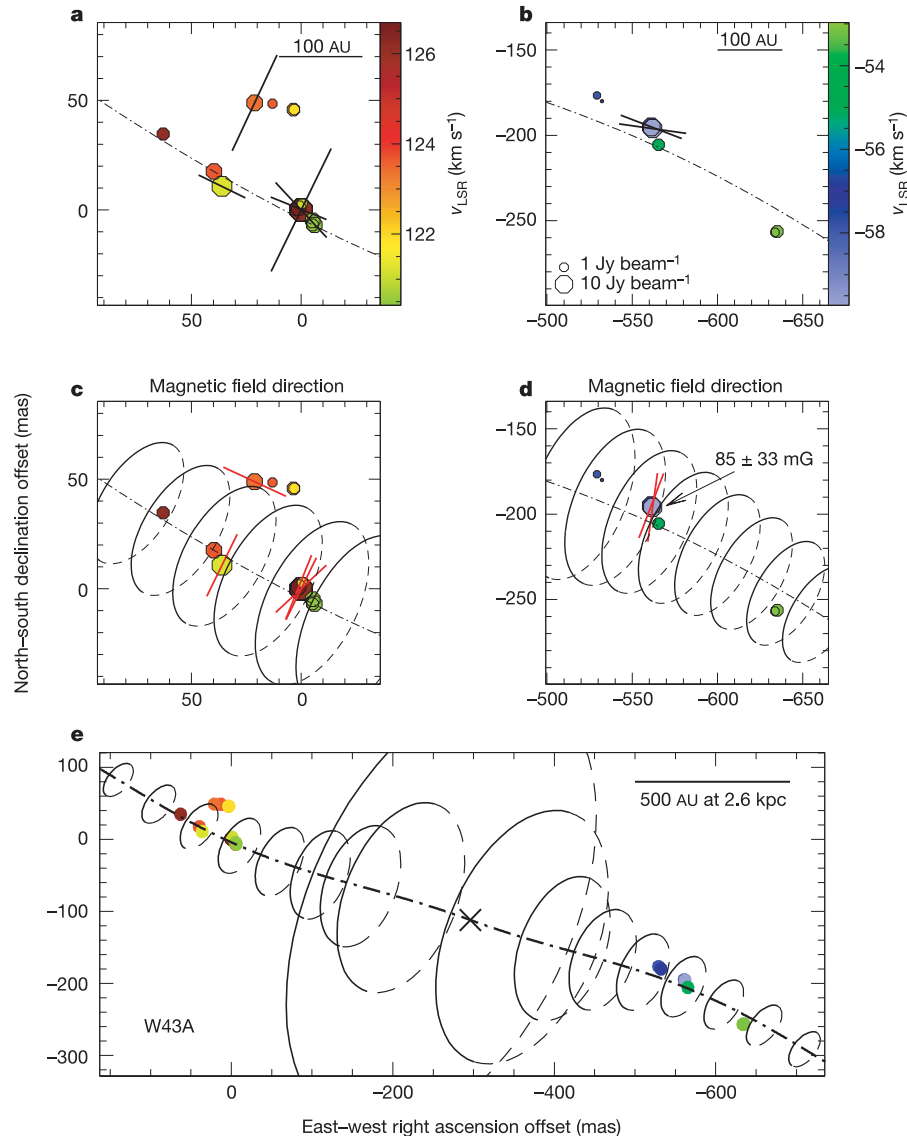


Figure 1 | The spatial distribution, linear polarization vectors and inferred magnetic field direction of the H₂O masers in the jet of W43A. H₂O maser emission (6₁₆–5₂₃ line at 22.235080 GHz) was observed with the VLBA with an angular resolution of 0.5 milliarcseconds (mas) in right ascension (α) and 1.0 mas in declination (δ). In **a–d**, the maser features are indicated by coloured hexagons, scaled logarithmically according to their flux density in Jy beam⁻¹. The maser features in **e** are indicated by coloured dots. The colour denotes the radial velocity v_{LSR} with respect to the local standard of rest (LSR). The dashed-dotted line indicates the model of the precessing jet of W43A (outflow velocity $v = 145 \text{ km s}^{-1}$; inclination 39° ; position angle 65° east of north, 5° precession with a 55 yr period). The cross in **e** denotes the location of W43A at approximately²⁷ $\alpha(\text{J2000}) = 18 \text{ h } 47 \text{ min } 41.166 \text{ s}$ and $\delta(\text{J2000}) = -01^\circ 45' 11.7''$. In **a** and **b**, the logarithmically scaled linear

polarization vectors are shown. In **c** and **d**, we indicate the derived magnetic field direction with red vectors. The vectors reveal a toroidal magnetic field along the jet. The perpendicular magnetic field direction of the northernmost maser feature indicates that these masers are probably located at the edge of the jet, where the magnetic field direction is along the projected jet direction owing to shear between the compressed toroidal field lines inside, and the uncompressed field lines outside, the jet. This indicates a $\sim 5^\circ$ jet opening angle. The magnetic field strength derived from the circular polarization (Fig. 2) is indicated in **d**. In **c–e**, the solid/dashed ellipses show the toroidal magnetic field component along the jet. The sizes of the ellipses scale according to a r^{-1} decrease of the magnetic field strength with increasing distance r from the star.

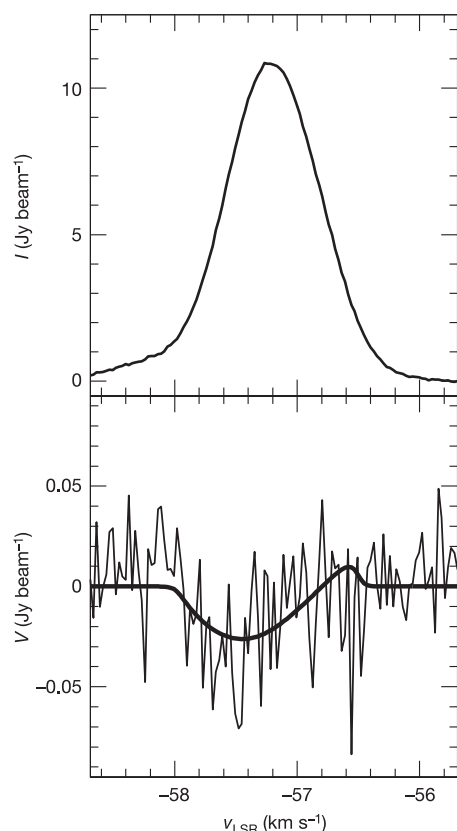


Figure 2 | The total power (I) and circular polarization spectrum (V) of a 22 GHz H_2O maser feature in the southern tip of the collimated jet of W43A. The spectra have a velocity resolution of 0.027 km s^{-1} and a r.m.s. noise of 14 mJy beam^{-1} . The feature has a peak flux of $10.85 \text{ Jy beam}^{-1}$ and a radial velocity $v_{\text{LSR}} = -57.19 \text{ km s}^{-1}$. The thick solid line in the lower panel is the best fit model to the circular polarization spectrum giving a circular polarization fraction $P_V = 0.33 \pm 0.09\%$. The s.d. error is determined from the r.m.s. noise on the total intensity and circular polarization spectra. The resulting magnetic field strength on the maser feature is $B_{\parallel} = 85 \pm 33 \text{ mG}$, where the increased s.d. error is the result of an added uncertainty due to the H_2O maser radiative transfer model used in the fit.

HH212 (ref. 11) and a large sample of AGN⁸. Additionally, recent laboratory work has been able to produce jets collimated by toroidal magnetic fields²⁹. For the jets produced by AGN there are many indications that magnetic collimation is occurring^{8,9,30}, however, no observations yet exist of the magnetic field strength. There is also no direct evidence of magnetic collimation of proto-stellar jets. The characteristics of the jet of W43A are similar to those found in star-forming environments¹¹, and as magnetic fields of similar strengths are found in those regions¹³, magnetic collimation such as described here probably also occurs during star formation.

Received 12 October; accepted 22 November 2005.

1. Sahai, R. & Trager, J. T. Multipolar bubbles and jets in low-excitation planetary nebulae: toward a new understanding of the formation and shaping of planetary nebulae. *Astron. J.* **116**, 1357–1366 (1998).
2. Imai, H., Obara, K., Diamond, P. J., Omodaka, T. & Sasao, T. A collimated jet of molecular gas from a star on the asymptotic giant branch. *Nature* **417**, 829–831 (2002).
3. Miranda, L. F., Gómez, Y., Anglada, G. & Torrelles, J. M. Water-maser emission from a planetary nebula with a magnetized torus. *Nature* **414**, 284–286 (2001).

4. Blackman, E. G., Frank, A., Markiel, J. A., Thomas, J. H. & Van Horn, H. M. Dynamos in asymptotic-giant-branch stars as the origin of magnetic fields shaping planetary nebulae. *Nature* **409**, 485–487 (2001).
5. García-Segura, G., López, J. A. & Franco, J. Magnetically driven winds from post-asymptotic giant branch stars: solutions for high-speed winds and extreme collimation. *Astrophys. J.* **618**, 919–925 (2005).
6. Pearson, T. J. & Readhead, A. C. S. The milliarcsecond structure of a complete sample of radio sources. II—First-epoch maps at 5 GHz. *Astrophys. J.* **328**, 114–142 (1988).
7. Blandford, R. D. & Payne, D. G. Hydromagnetic flows from accretion discs and the production of radio jets. *Mon. Not. R. Astron. Soc.* **199**, 883–903 (1982).
8. Xu, C., Livio, M. & Baum, S. Radio-loud and radio-quiet Active Galactic Nuclei. *Astron. J.* **118**, 1169–1176 (1999).
9. Gabuzda, D. C., Murray, É. & Cronin, P. Helical magnetic fields associated with the relativistic jets of four BL Lac objects. *Mon. Not. R. Astron. Soc.* **351**, L89–L93 (2004).
10. Ouyed, R., Pudritz, R. E. & Stone, J. M. Episodic jets from black holes and protostars. *Nature* **385**, 409–414 (1997).
11. Zinnecker, H., McCaughrean, M. J. & Rayner, J. T. A symmetrically pulsed jet of gas from an invisible protostar in Orion. *Nature* **394**, 862–865 (1998).
12. Diamond, P. J., Norris, R. P., Rowland, P. R., Booth, R. S. & Nyman, L.-A. The circumstellar envelopes around OH/IR stars. *Mon. Not. R. Astron. Soc.* **212**, 1–21 (1985).
13. Vlemmings, W. H. T., Diamond, P. J., van Langevelde, H. J. & Torrelles, J. M. The magnetic field in the star-forming region Cepheus A from H_2O maser polarization observations. *Astron. Astrophys.* (in the press); preprint at (<http://arxiv.org/astro-ph/0510452>) (2005).
14. Vlemmings, W. H. T., Diamond, P. J. & van Langevelde, H. J. Circular polarization of water masers in the circumstellar envelopes of late type stars. *Astron. Astrophys.* **394**, 589–602 (2002).
15. Vlemmings, W. H. T., van Langevelde, H. J. & Diamond, P. J. The magnetic field around late-type stars revealed by the circumstellar H_2O masers. *Astron. Astrophys.* **434**, 1029–1038 (2005).
16. Nedoluha, G. E. & Watson, W. D. The Zeeman effect in astrophysical water masers and the observation of strong magnetic fields in regions of star formation. *Astrophys. J.* **384**, 185–196 (1992).
17. Goldreich, P., Keeley, D. A. & Kwan, J. Y. Astrophysical masers. II. Polarization properties. *Astrophys. J.* **179**, 111–134 (1973).
18. Kemball, A. J. & Diamond, P. J. Imaging the magnetic field in the atmosphere of TX Cam. *Astrophys. J.* **481**, L111–L114 (1997).
19. Elitzur, M. *Astronomical Masers* (Kluwer Academic, Dordrecht, 1992).
20. Crutcher, R. M. Magnetic fields in molecular clouds: observations confront theory. *Astrophys. J.* **520**, 706–713 (1999).
21. Elitzur, M., Hollenbach, D. J. & McKee, C. F. Planar H_2O masers in star-forming regions. *Astrophys. J.* **394**, 221–227 (1992).
22. Reid, M. J. et al. Stellar OH masers and magnetic fields—VLBI observations of U Orionis and IRC +10420. *Astrophys. J.* **227**, L89–L92 (1979).
23. Chevalier, R. A. & Luo, D. Magnetic shaping of planetary nebulae and other stellar wind bubbles. *Astrophys. J.* **421**, 225–235 (1994).
24. García-Segura, G., Langer, N., Różyczka, M. & Franco, J. Shaping bipolar and elliptical planetary nebulae: effects of stellar rotation, photoionization heating, and magnetic fields. *Astrophys. J.* **517**, 767–781 (1999).
25. Sahai, R. Bipolar and multipolar jets in protoplanetary and planetary nebulae. *Rev. Mex. Astron. Astrofis.* **13**, 133–138 (2002).
26. Blackman, E. G. in *Asymmetrical Planetary Nebulae III: Winds, Structure and the Thunderbird* (eds Meixner, M., Kastner, J. H., Balick, B. & Soker, N.) 401–407 (ASP Conference Series Vol. 313, Astronomical Society of the Pacific, San Francisco, 2004).
27. Imai, H., Nakashima, J., Diamond, P. J., Miyazaki, A. & Deguchi, S. A biconically expanding flow in W43A traced by SiO maser emission. *Astrophys. J.* **622**, L125–L128 (2005).
28. Lery, T. et al. Magnetised protostellar bipolar outflows. I. Self-similar model with Poynting flux. *Astron. Astrophys.* **350**, 254–274 (1999).
29. Lebedev, S. V. et al. Magnetic tower outflows from a radial wire array Z-pinch. *Mon. Not. R. Astron. Soc.* **361**, 97–108 (2005).
30. Asada, K. et al. A helical magnetic field in the jet of 3C 273. *Publ. Astron. Soc. Jpn* **54**, L39–L43 (2002).

Acknowledgements NRAO is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. W.H.T.V. was supported by a Marie-Curie Intra-European fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.H.T.V. (wouter@jb.man.ac.uk).

Episodic outgassing as the origin of atmospheric methane on Titan

Gabriel Tobie^{1,2}, Jonathan I. Lunine^{2,3} & Christophe Sotin¹

Saturn's largest satellite, Titan, has a massive nitrogen atmosphere containing up to 5 per cent methane near its surface. Photochemistry in the stratosphere would remove the present-day atmospheric methane in a few tens of millions of years¹. Before the Cassini-Huygens mission arrived at Saturn, widespread liquid methane or mixed hydrocarbon seas hundreds of metres in thickness were proposed as reservoirs from which methane could be resupplied to the atmosphere over geologic time². Titan fly-by observations^{3–5} and ground-based observations⁶ rule out the presence of extensive bodies of liquid hydrocarbons at present, which means that methane must be derived from another source over Titan's history. Here we show that episodic outgassing of methane stored as clathrate hydrates within an icy shell above an ammonia-enriched water ocean is the most likely explanation for Titan's atmospheric methane. The other possible explanations all fail because they cannot explain the absence of surface liquid reservoirs and/or the low dissipative state of the interior. On the basis of our models, we predict that future fly-bys should reveal the existence of both a subsurface water ocean and a rocky core, and should detect more cryovolcanic edifices.

The presence of methane within the atmosphere and the interior is primarily determined by the nature of the materials from which Titan was built, and the bulk density of Titan indicates that the body is roughly 0.5–0.7 by mass silicate (for a silicate density of $3,000\text{--}4,000\text{ kg m}^{-3}$) and the remainder is mostly water ice^{7,8}. Even though external sources of methane derived from cometary impact may have been significant in a far past, they cannot explain its current abundance⁹. Furthermore, direct sampling of Titan's atmosphere by the mass spectrometers on board the Orbiter¹⁰ and Probe¹¹ provide circumstantial evidence that ammonia (NH_3) is the primordial source of Titan's atmospheric molecular nitrogen⁹. For ammonia to be the source of the nitrogen, the hydrogen must have been removed by atmospheric photochemistry, charged particle chemistry or impacts, so that part of the ammonia brought by the accreted materials should have been devolatilized during or soon after Titan's formation. This is consistent with previous formation models^{12–14}, and a consistent post-accretion internal structure comprises an inner homogeneous core made of a mixture of rock and icy materials, overlain by a silicate layer and a water layer enriched with liquid ammonia¹² (see Fig. 2c).

At the end of accretion, only the inner undifferentiated portion of Titan's interior was able to hold methane. As most of the region outward of this proto-core was warm liquid water, in which methane has very low solubility, potentially very large amounts of methane ended up in the primitive atmosphere and on the surface. Methane in the inner core could have been incorporated there during accretion, reflecting its presence in the planetesimal building blocks of Titan, or it might have come from the conversion of CO_2 through 'serpentinization' in the silicate-rich core in a way similar to the process

involved in the Earth's hydrothermal systems¹⁵. Whatever its origin, laboratory experiments^{16,17} show that methane under the temperature–pressure conditions of Titan's cold, undifferentiated proto-core would rapidly combine with water molecules to form methane clathrate hydrate¹⁷. Owing to its low density (see Supplementary Information 1.1), any methane clathrate released during the overturn that formed a discrete rock core¹² would ascend to the top of the outer liquid layer without being dissociated because of its high stability¹⁸ (see Supplementary Information 1.2).

Subsequent to core overturn, methane outgassing would have occurred only when the conditions required to dissociate methane clathrate were reached within the icy mantle of Titan's interior. We use a coupled thermal–orbital model¹⁹ to determine the conditions and timing of clathrate dissociation. Although some of the latest interior models^{18,20} indicate that methane clathrate dissociation within Titan's interior is very difficult, it must be noted that those models included neither the coupling between Titan's interior and its orbit nor the low thermal conductivity¹⁷ and high viscosity²¹ of methane clathrate as compared with water ice. Here we show that the incorporation of those effects strongly influences the cooling rate of Titan's interior and leads to episodic outgassings of methane. Our

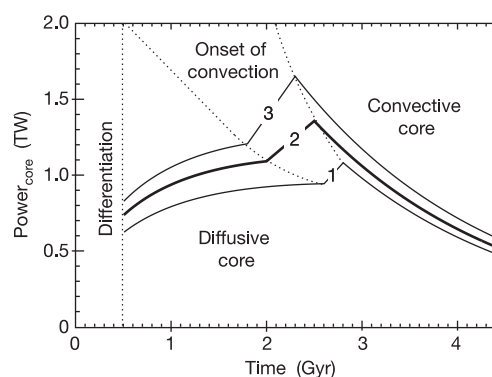


Figure 1 | Global heat power provided by the rocky core. Three different mineral compositions, reflecting different hydration states, have been used to estimate the possible evolution of the silicate core over a reasonable range: 1, $\rho_{\text{sil}} = 3,000\text{ kg m}^{-3}$, $H_0 = 3 \times 10^{-11}\text{ W kg}^{-1}$, $T_i = 1,400\text{ K}$; 2, $\rho_{\text{sil}} = 3,300\text{ kg m}^{-3}$, $H_0 = 4 \times 10^{-11}\text{ W kg}^{-1}$, $T_i = 1,500\text{ K}$; 3, $\rho_{\text{sil}} = 4,000\text{ kg m}^{-3}$, $H_0 = 5 \times 10^{-11}\text{ W kg}^{-1}$, $T_i = 1,600\text{ K}$, where ρ_{sil} is the silicate density, H_0 is the initial radiogenic heating rate and T_i is the interior temperature required to trigger thermal convection in the silicate core. Evolution of the silicate core is characterized by a diffusive stage following the differentiation and a convective stage once the interior is sufficiently hot. Depending on the composition of the silicate core, we estimate an onset time of convection of between 1.8 and 2.8 Gyr (see Supplementary Information 2.1).

¹Laboratoire de Planétologie et Géodynamique, UMR-CNRS 6112, Université de Nantes, 44322 Nantes cedex 03, France. ²Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. ³Instituto di Fisica dello Spazio Interplanetario, INAF, Via del Fosso del Cavaliere 100, 00133 Roma, Italia.

model includes heat transfer through the outer layer, clathrate dissociation and crystallization of the liquid layer, thermal evolution of the silicate core, and tidal dissipation and its effect on orbital eccentricity decay (see Supplementary Information 2).

Titan shows an orbit with an eccentricity of 3%, which cannot be explained by orbital resonance. Titan's current free eccentricity is probably the fossil of a higher primordial eccentricity¹⁹. We therefore consider a wide range of initial eccentricity values (5–30%), and we retain only models that reproduce the current value. We also explore different interior compositions with varying rheological assumptions and different concentrations of radiogenic elements. Figure 1 shows a reasonable range of global heat power coming out of the silicate core, assuming silicate minerals with different compositions.

A typical example of interior evolution and methane outgassing is shown in Fig. 2. Three major episodes of outgassing are predicted and result from, first, internal differentiation (silicate core formation); second, onset of convection in the silicate core; and third, late subsequent cooling and crystallization of the outer layer. The first episode follows the overturn of the undifferentiated proto-core with the silicate carapace and the simultaneous formation of the methane clathrate layer above the ocean. The clathrate accumulation near the surface results in an increase in the ocean temperature, owing to the low conductivity and the high viscosity of the methane clathrate, and hence release of methane. The second episode occurs when the

silicate core is sufficiently hot for thermal convection to initiate. The heat flux at the surface of the core increases, which leads to a thinner outer clathrate-rich layer. Methane accumulates at the base of the outer layer because the ocean is saturated. The methane can ascend along buoyancy-driven cracks¹². Afterwards, the silicate core starts cooling down. The third and last episode is a natural consequence of both the interior cooling and the crystallization of the liquid layer. Once the ice I layer that crystallizes below the clathrate layer becomes unstable against thermal convection (at 3.9 gigayears (Gyr) in Fig. 2), hot thermal plumes, favoured by tidal dissipation²², penetrate the clathrate layer and induce clathrate dissociation. As shown by Figs 2a and 3, the final outgassing episode is more than enough to maintain the atmospheric abundance at or above the current level of 5% (ref. 11), and indeed to saturate the atmosphere and hence soak the surface in extensive methane rainfall—evidence of which is seen in the data^{4,23}. Our calculations show that the final episode must have started relatively late (after 3.5 Gyr), to explain the high value of orbital eccentricity¹⁹ and to induce a sufficiently high outgassing rate (Fig. 3).

Our numerical results provide an explanation for the appearance of flow features and volcanic edifices, as suggested by both infrared³ and radar⁴ observations, and circumstantial evidence of shallow subsurface reservoirs of methane^{4,11,23}. Although methane outgassing could have occurred in the past (before 2.5 Gyr), only recent outgassing can

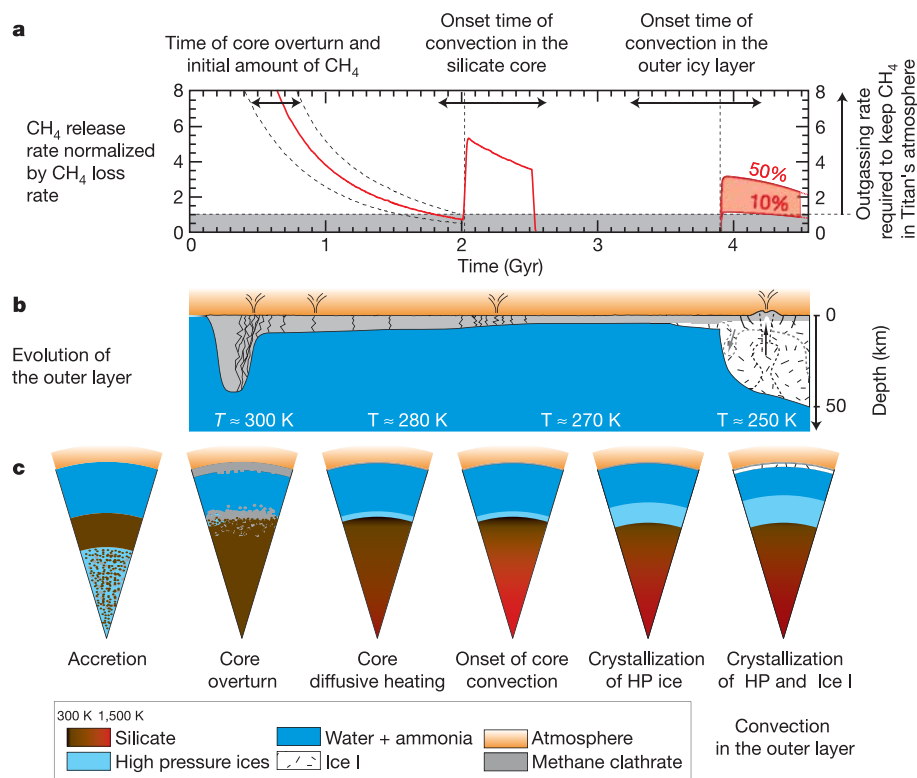


Figure 2 | The evolution of the interior and the outgassing rate of methane over Titan's history. In the simulation shown, we use composition 2 in Fig. 1 for evolution of the silicate core and we assume an ammonia fraction of 5% and a value of 10^{14} Pa s for the viscosity of ice close to its melting point. Furthermore, the initial eccentricity is set to 13.5% to fit the current value. We start the evolution at time, $t = 0.5$ Gyr once the formation of the silicate core is achieved¹² and once the energy contribution from meteoritic bombardment can be safely neglected. The methane release rate, normalized by the current loss rate ($1.25 \times 10^{10} \text{ cm}^{-2} \text{ s}^{-1}$; **a**), is controlled by the interior evolution (**c**) and especially the outer region evolution (**b**) and is characterized by three major outgassing episodes. The amount of methane outgassed during the two first episodes is determined mainly by the initial amount of methane in the proto-core, the time and duration of the core overturn (~ 0.2 – 1 Gyr), and the efficiency of convection in the silicate core.

The third and last episode is triggered by upwelling icy plumes, whose typical temperature (~ 250 K) is close to the dissociation temperature of methane clathrate hydrate at a depth of 1 km. The maximum outgassing rate is expected just after the onset of convection (~ 4 Gyr), when the surface heat flow is maximal. Then, the methane release rate is expected to diminish as tidal dissipation decreases owing to eccentricity decay and as the outer layer and the conductive lid thicken. Panel **a** shows a representative methane outgassing rate from 3.9 to 4.55 Gyr, assuming the decomposition of 10–50% of the clathrate mass remaining in the crust before the onset of convection. Arrows indicate the possible onset time of outgassing obtained for varying internal composition and rheology. We note that, for the large range of parameter values explored, the three episodes of outgassing may differ only in duration and vigour.

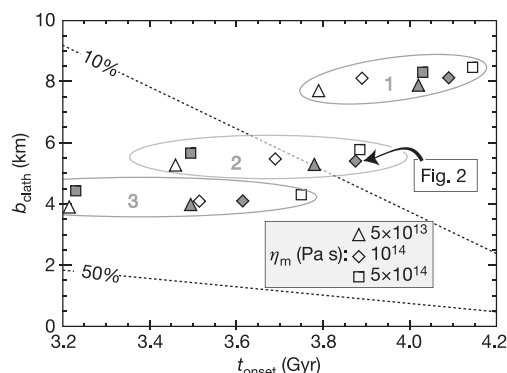


Figure 3 | Thickness of the available methane clathrate reservoir when the final episode initiates. The beginning of the third episode is related to the onset time (t_{onset}) of thermal convection in the outer layer, and depends mainly on the ammonia fraction x_{NH_3} and the viscosity of ice I near its melting point η_m . Filled symbols, $x_{\text{NH}_3} = 5\%$; open symbols, $x_{\text{NH}_3} = 0\%$. Labels 1, 2 and 3 refer to the three different schemes for evolution of the silicate core shown in Fig. 1. The two dotted lines represent the reservoir thickness required to counterbalance the methane loss rate ($1.25 \times 10^{10} \text{ cm}^{-2} \text{ s}^{-1}$), assuming that 10 or 50% of the clathrate reservoir has been outgassed since the onset of the last episode. We show that the thickness of the available reservoir (b_{clath}) is determined mainly by the evolution of the silicate core and that the duration and vigour of the last episode depends on the ammonia fraction x_{NH_3} and the ice viscosity at the melting point η_m . Even though a third outgassing episode starting as early as 3–3.5 Gyr is possible for some internal parameters, it requires either initial eccentricities $>20\%$, which is difficult to explain in view of the proximity of Hyperion¹⁹, or a high-viscosity value ($5 \times 10^{14} \text{ Pa s}$) and an ammonia fraction of 5%, which leads to low surface heat flow ($<10 \text{ mW m}^{-2}$) and hence a very low outgassing rate. Nevertheless, our calculations show that models with some fraction of ammonia and a relatively low concentration of radiogenic elements in the silicate (core schemes 1 and 2) lead to both a late onset ($t > 3.8 \text{ Gyr}$) and a high outgassing rate. Such models require lower initial eccentricity values ($<10\text{--}15\%$) and retain larger masses of clathrate reservoir through Titan's history.

explain the presence of methane in the current atmosphere in the absence of very massive, and hence surface-exposed, reservoirs to buffer the loss of atmospheric methane in a short time ($\sim 10^7 \text{ yr}$) as compared with the time elapsed since outgassing episodes 1 and 2. The third outgassing episode would be due to thermal destabilization of the clathrate layer to form a methane-filled chamber, which would be outgassed periodically or catastrophically through eruptive processes. Our finding is fully consistent with the carbon isotopic ratio $^{13}\text{C}/^{12}\text{C}$ in hydrocarbon molecules close to the terrestrial value^{11,24}. This ratio reflects a bulk carbon inventory that did not participate in the massive escape phase of the ancient post-accretionary atmosphere, in contrast to nitrogen, whose isotopic ratio $^{15}\text{N}/^{14}\text{N}$ is enhanced and thus suggests escape^{11,24}.

The presence of a significant amount in the atmosphere of the ^{40}K decay daughter ^{40}Ar (refs 10, 11) (10% of the maximum amount that could possibly be outgassed assuming a chondritic composition for the rock core) can be understood in the framework of our model. The interior was warm enough in the past (before 2.5 Gyr) for the water ocean to be enriched in potassium through interactions with the rocky core²⁵. Argon could then have ended up in the atmosphere if potassium-enriched water cryomagmas were directly deposited on the surface²⁵ or, more likely, if it was outgassed simultaneously with methane. Although argon incorporates in the clathrate hydrate to a lesser extent than methane^{26,27}, mobilization of argon contained in the water ocean through methane bubble formation (occurring during episodes 1 and 2), as well as successive destabilization of methane clathrate containing traces of argon, may progressively enrich the atmosphere in argon. The presence of argon is thus a strong indicator of past and recent internal activities.

Future fly-by observations will provide pertinent tests of our ideas. If a subsurface ammonia–water ocean and a rocky core are present, as predicted by our model, Titan's tidal secondary potential Love number k_2 should be higher than 0.3 (ref. 28) and its moment of inertia factor C/MR^2 should be lower than 0.33; these quantities should be measured by the Radio Science Subsystem later in the mission²⁹. Our model predicts that a major outgassing event occurred about 2 billion years ago, followed by cessation of methane injection and associated cryovolcanic activity until a smaller episode in the recent past. This has two testable consequences.

First, the cratering record should show a surface age neither significantly more nor significantly less than 2 billion years. Since the two impact craters found so far⁴ imply a slightly smaller age if no further such structures exist, we predict that several more large (50–100 km) impact craters—intact or partially eroded/buried—will be found by the Orbiter. Moreover, depletion of methane before the final outgassing episode may have cooled and thinned the atmosphere³⁰, augmenting the number of smaller impact craters relative to what is expected with the screening effects of today's atmosphere. Second, the two first episodes would have been widespread and able to put a large amount of liquid methane on the surface for extended periods of time, whereas the recent one would be more localized and controlled by the activity of hot ice plumes. Evidence for modification of large areas of the surface by the action of liquid methane seems to be present in the radar data in the form of widespread fluvial channels and relatively featureless (sediment-filled?) basins⁴; additional remote sensing observations and crater-density statistics should reveal geological differences between terrains formed during the ancient widespread episodes and recently modified terrains associated with cryovolcanic edifices.

Received 22 June; accepted 28 November 2005.

- Yung, Y. L., Allen, M. A. & Pinto, J. P. Photochemistry of the atmosphere of Titan: comparison between model and observations. *Astrophys. J. Suppl. Ser.* **55**, 465–506 (1984).
- Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
- Sotin, C. et al. Release of volatiles from a possible cryovolcano from near-infrared imaging of Titan. *Nature* **435**, 786–789 (2005).
- Elachi, C. et al. Cassini radar views the surface of Titan. *Science* **308**, 970–974 (2005).
- Porco, C. et al. Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
- West, R. A., Brown, M. E., Salinas, S. V., Bouchez, A. H. & Roe, H. G. No oceans on Titan from the absence of a near-infrared specular reflection. *Nature* **436**, 670–672 (2005).
- Stevenson, D. J. in *Proceedings Symposium on Titan, Toulouse, France, 9–12 September 1991* 29–33 (European Space Agency, Noordwijk, The Netherlands, 1992).
- Grasset, O., Sotin, C. & Deschamps, F. On the internal structure and dynamics of Titan. *Planet. Space Sci.* **48**, 617–636 (2000).
- Owen, T. On the origin of Titan's atmosphere. *Planet. Space Sci.* **48**, 747–752 (2000).
- Waite, H. et al. Ion-neutral mass spectrometer results from the first flyby of Titan. *Science* **308**, 982–986 (2005).
- Niemann, H. et al. The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* **438**, 779–784 (2005).
- Lunine, J. I. & Stevenson, D. J. Clathrate and ammonia hydrates at high pressure—application to the origin of methane on Titan. *Icarus* **70**, 61–77 (1987).
- Kuramoto, K. & Matsui, T. Formation of a hot proto-atmosphere on the accreting giant icy satellite: Implications for the origin and evolution of Titan, Ganymede, and Callisto. *J. Geophys. Res.* **99**, 21183–21200 (1994).
- Mousis, O., Gautier, D. & Bockelée-Morvan, D. An evolutionary turbulent model of Saturn's subnebula: implications for the origin of the atmosphere of Titan. *Icarus* **156**, 162–175 (2002).
- Lowell, R. P. & Rona, P. A. Seafloor hydrothermal systems driven by the serpentinization of peridotite. *Geophys. Res. Lett.* **29**, 1531, doi:10.1029/2001GL014411 (2002).
- Loveday, J. S. et al. Stable methane hydrate above 2 GPa and the source of Titan's atmospheric methane. *Nature* **410**, 661–663 (2001).
- Sloan, E. D. *Clathrates Hydrates of Natural Gases* 2nd edn (Marcel Dekker, New York, 1998).

18. Grasset, O. & Pargamin, J. The ammonia water system at high pressures: implications for the methane of Titan. *Planet. Space Sci.* **53**, 371–384 (2005).
19. Tobie, G., Grasset, O., Lunine, J. I., Mocquet, A. & Sotin, C. Titan's internal structure inferred from a coupled thermal-orbital model. *Icarus* **175**, 496–502 (2005).
20. Sohl, F., Hussmann, H., Schwentker, B., Spohn, T. & Lorenz, R. D. Interior structure models and tidal Love numbers of Titan. *J. Geophys. Res.* **108**, 5130, doi:10.1029/2003JE002044 (2003).
21. Durham, W. B., Kirby, S. H., Stern, L. A. & Zhang, W. The strength and rheology of methane clathrate hydrate. *J. Geophys. Res.* **108**, 2182, doi:10.1029/2002JB001872 (2003).
22. Sotin, C., Head, J. W. & Tobie, G. Europa: tidal heating of upwelling thermal plumes and the origin of lenticulae and chaos melting. *Geophys. Res. Lett.* **29**, 1233, doi:10.1029/2001GL013844 (2002).
23. Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* **438**, 765–778 (2005).
24. Lunine, J. I., Yung, Y. L. & Lorenz, R. D. On the volatile inventory of Titan from isotopic abundances in nitrogen and methane. *Planet. Space Sci.* **47**, 1291–1303 (1999).
25. Engel, S., Lunine, J. I. & Norton, D. L. Silicate interactions with ammonia-water fluids on early Titan. *J. Geophys. Res.* **99**, 3745–3752 (1994).
26. Dyadin, Y. A., Aladko, E. Y. & Larionov, E. G. Decomposition of methane hydrates up to 15 kbar. *Mendeleev Commun.* **7**, 34–35 (1997).
27. Dyadin, Y. A., Larionov, E. G., Mirinski, D. S., Mikina, T. V. & Starostina, L. I. Clathrate formation in the Ar–H₂O system under pressures up to 15000 bar. *Mendeleev Commun.* **7**, 32–34 (1997).
28. Castillo, J., Rappaport, J., Moquet, A. & Sotin, C. Clues on Titan's internal structure from Cassini-Huygens mission. *Lunar Planet. Sci.* (LPI contribution no. 1109) **33**, 1989 (2002).
29. Rappaport, N., Bertotti, B., Giamperi, G. & Anderson, J. D. Doppler measurements of the quadrupole moments of Titan. *Icarus* **126**, 313–323 (1997).
30. Lorenz, R., McKay, C. P. & Lunine, J. I. Photochemical collapse of Titan's atmosphere. *Science* **275**, 642–646 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The first author benefits from post-doctoral fellowships from the Lavoisier programme of the 'Ministères des Affaires étrangères' and from the Centre National d'Etudes Spatiales (France). Support for the project and paper was provided by the Cassini Project and by the INSU 'Programme National de Planétologie'.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.T. (gabriel.tobie@univ-nantes.fr).

Hidden magnetism and quantum criticality in the heavy fermion superconductor CeRhIn₅

Tuson Park¹, F. Ronning¹, H. Q. Yuan², M. B. Salamon², R. Movshovich¹, J. L. Sarrao¹ & J. D. Thompson¹

With only a few exceptions that are well understood, conventional superconductivity does not coexist with long-range magnetic order (for example, ref. 1). Unconventional superconductivity, on the other hand, develops near a phase boundary separating magnetically ordered and magnetically disordered phases^{2,3}. A maximum in the superconducting transition temperature T_c develops where this boundary extrapolates to zero Kelvin, suggesting that fluctuations associated with this magnetic quantum-critical point are essential for unconventional superconductivity^{4,5}. Invariably, though, unconventional superconductivity masks the magnetic phase boundary when $T < T_c$, preventing proof of a magnetic quantum-critical point⁵. Here we report specific-heat measurements of the pressure-tuned unconventional superconductor CeRhIn₅ in which we find a line of quantum-phase transitions induced inside the superconducting state by an applied magnetic field. This quantum-critical line separates a phase of coexisting antiferromagnetism and superconductivity from a purely unconventional superconducting phase, and terminates at a quantum tetracritical point where the magnetic field completely suppresses superconductivity. The $T \rightarrow 0$ K magnetic field–pressure phase diagram of CeRhIn₅ is well described with a theoretical model^{6,7} developed to explain field-induced magnetism in the high- T_c copper oxides, but in which a clear delineation of quantum-phase boundaries has not been possible. These experiments establish a common relationship among hidden magnetism, quantum criticality and unconventional superconductivity in copper oxides and heavy-electron systems such as CeRhIn₅.

CeRhIn₅ belongs to a family of cerium-, uranium- and plutonium-based compounds in which interactions between electrons enhance the effective mass of charge carriers up to 1,000 times the mass of a free electron. These heavy-electron materials in turn belong to a larger family of strongly correlated electron systems that includes the high- T_c copper oxide superconductors and some organic compounds. A generic temperature–control parameter (T – δ) phase diagram common to strongly correlated unconventional superconductors is shown in Fig. 1 (refs 5, 8–10). CeRhIn₅ is prototypical of this phase diagram, in this case with pressure as the tuning parameter^{11,12}. At atmospheric pressure, CeRhIn₅ orders antiferromagnetically at 3.8 K, and with applied pressure the antiferromagnetic (AFM) state vanishes at $P_{c1} = 1.77$ GPa when the Néel temperature T_N equals the superconducting transition temperature $T_c = 1.9$ K. Over a range of pressures below P_{c1} , extensive measurements show^{12–15} that magnetic order coexists with superconductivity, but only when $T_N > T_c$. Above P_{c1} , where $T_c > T_N$ (extrapolated), these measurements find only unconventional superconductivity. A smooth extrapolation of $T_N(P)$ to $T = 0$ K suggests that Néel order, if it existed, would terminate at a quantum critical point P_{c2} near 2.3 GPa, where the effective mass of charge carriers diverge in the normal state¹⁶.

Figure 2 summarizes field-dependent specific-heat measurements on a single crystal of CeRhIn₅ subjected to pressures just below and above P_{c1} . For $P = 1.68$ GPa $< P_{c1}$ (Fig. 2a), these data confirm earlier conclusions^{12–18} that an AFM order coexists with the superconducting phase. Above P_{c1} (Fig. 2b), only a specific-heat discontinuity due to superconductivity is observed for fields up to 44 kOe: there is no evidence for a magnetic phase transition at these low fields, consistent with other measurements at zero applied field^{12–15,19}. At 55 kOe, a specific-heat anomaly near 0.7 K emerges below the superconducting transition ($T_c = 1.7$ K) and grows in intensity with increasing field. Finally, in Fig. 2c ($P = 2.3$ GPa), there is no signature for a magnetic transition up to 88 kOe and down to 300 mK. As the superconducting transition is suppressed to zero, the ratio between

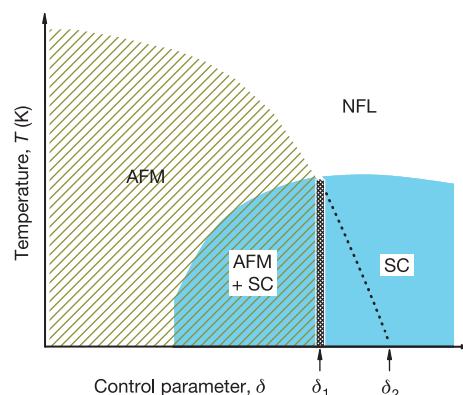


Figure 1 | Schematic temperature–control parameter (T – δ) phase diagram common to classes of unconventional superconductors. The hatched area represents an AFM state and the coloured area represents a superconducting (SC) phase. The hatched area in the coloured background denotes coexisting AFM and SC phases. In the normal state above the SC dome, physical properties are not typical of a metal and reflect non-Fermi liquid (NFL) behaviours. As the control parameter, such as chemical substitution or pressure is varied, long-range magnetic order gives way to a superconducting state. Above the superconducting dome, normal state properties are dominated by long-ranged, long-lived fluctuations that are expected if the magnetic phase boundary extended smoothly to absolute zero temperature, that is, to a magnetic quantum critical point (δ_2). Experimentally, however, magnetic order abruptly disappears at a finite temperature where the superconductivity and magnetic phase boundaries meet, suggesting a first-order or weakly first-order boundary at δ_1 and providing no obvious connection between magnetism and the putative δ_2 . In such a case, it is difficult to reconcile the existence of an extended range of unconventional superconductivity beyond δ_1 and of an unusual normal state above T_c .

¹Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. ²Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

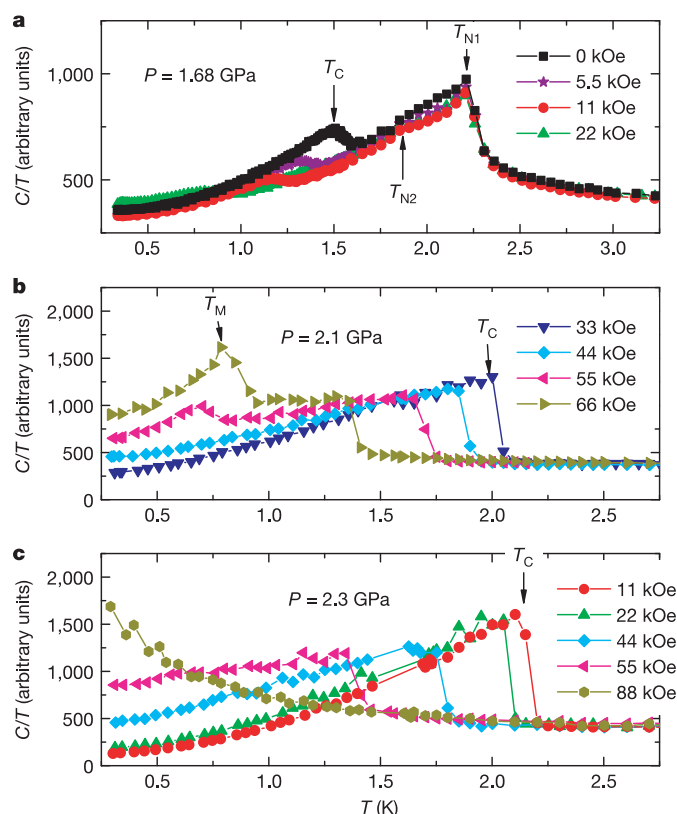


Figure 2 | Specific heat divided by temperature as a function of temperature for CeRhIn₅ at fixed magnetic fields. The magnetic field is applied perpendicular to the *c* axis of this tetragonal compound. Specific heat is determined by an alternating-current calorimetric method (see Supplementary Methods). **a**, At 1.68 GPa, specific-heat anomalies due to AFM transitions T_{N1} and T_{N2} and a superconducting transition T_c are observed. T_{N1} signals the onset of incommensurate antiferromagnetism with propagation wave vector (0.5, 0.5, 0.297) (ref. 17) and $T_{N2} = 1.85$ K is due to a spin reorientation transition¹⁸. As at atmospheric pressure, T_{N1} and T_{N2} are almost independent of the applied field. A specific-heat discontinuity due to superconductivity follows at 1.55 K. **b**, At 2.1 GPa, only a superconducting anomaly appears for $H \leq 44$ kOe (diamonds). At 55 kOe (side triangles), however, magnetism appears for $T < T_c(H)$. With further increasing field, the magnetic anomaly at T_M is enhanced and persists for $T > T_c(H)$, as shown in Supplementary Fig. 1. **c**, At 2.3 GPa, only superconductivity appears and magnetism is absent for $H \leq 88$ kOe and $T > 300$ mK.

specific heat and temperature C/T diverges weakly with decreasing temperature.

Figure 3a shows the evolution of the field-induced magnetic anomaly in C/T for $P = 2.1$ GPa. Similar results were obtained at $P = 1.8$ and 1.9 GPa, even closer to P_{c1} . The area under these curves is a measure of magnetic entropy (Fig. 3a inset) and reflects approximately the magnitude of the field-induced magnetism. The near-linear proportionality of the entropy to the applied field suggests that magnetism is associated with quantized vortices of magnetic flux that penetrate the superconductor and whose areal density is proportional to magnetic-field strength H . The H -induced transition temperature increases with H , consistent with the intrinsic magnetism, not with superconducting nor extrinsic phases. An explanation for these observations is discussed later.

A temperature–pressure phase diagram constructed from specific-heat measurements is plotted in Fig. 3b for representative fixed fields. For zero magnetic field, evidence for a magnetic transition temperature abruptly disappears at P_{c1} , suggesting a first-order-like transition. In an applied field of 33 kOe, however, the line of second-order magnetic transition temperatures T_{N1} smoothly

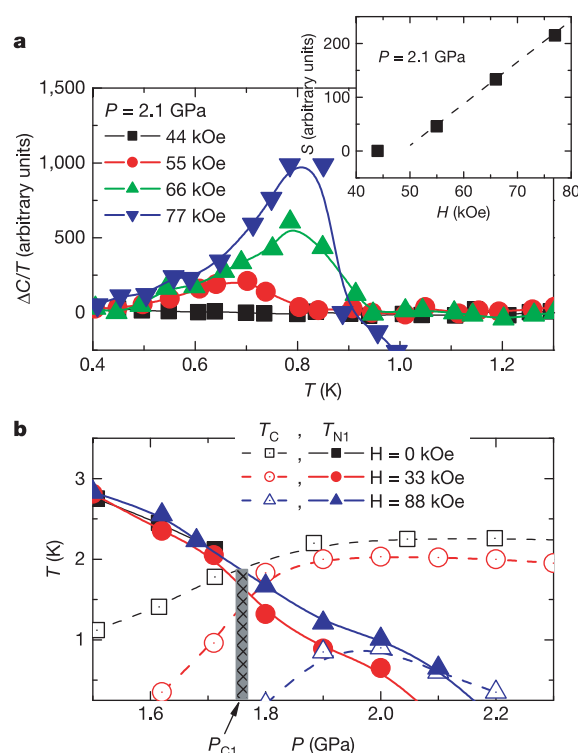


Figure 3 | Measurements of field and pressure dependence. **a**, Field-dependence of the magnetic specific heat of CeRhIn₅ at 2.1 GPa. To estimate the magnetic contribution to the specific heat $\Delta C/T$ due to magnetic order, we assume a smoothly varying background ($\sim T^2$) over a temperature range of interest and subtract that background contribution from the total measured specific heat. Inset, Entropy involved in the magnetic ordering transition, that is, area under the peak in $\Delta C/T$, as a function of magnetic field. The dashed line is a guide to eyes. The relative entropy associated with magnetic order is much less for $P = 2.1$ GPa than at $P = 0$ for any H . **b**, Temperature–pressure diagram at $H = 0$ kOe (squares), 33 kOe (circles), and 88 kOe (triangles). The magnetic transition is depicted as solid symbols and the superconducting transition by open symbols. Lines are spline fits to the data points.

evolves through P_{c1} and deep into the superconducting dome. With increasing field, the relative position of the critical points where T_{N1} becomes zero changes with respect to the centre of the superconducting dome. The influence of superconductivity on the development of magnetic order is reflected in a slope change of the magnetic transition line as it crosses into the superconducting domain.

These results are shown in Fig. 4 as a temperature–pressure–field phase diagram. The vertical field–pressure plane at a fixed temperature of 0.65 K changes very little with decreasing temperature to 350 mK, and we take it to be representative of a $T = 0$ K plane (see Supplementary Fig. 1). With increasing pressure, the magnetic field required to induce magnetism increases and finally meets the upper critical field line H_{c2} at $P_{c2} = 2.25$ GPa, a tetracritical point that branches out a transition line between magnetically ordered and disordered phases for $H > H_{c2}$. This observation of the H -induced magnetism only for $P_{c1} < P < P_{c2}$ now provides an explanation for the de Haas–van Alphen observation¹⁶ of a diverging effective mass due to a field- and pressure-tuned quantum critical point at P_{c2} . At this pressure, the Fermi surface volume expands to accommodate additional delocalized charge carriers. The larger Fermi surface volume of CeRhIn₅ beyond P_{c2} corresponds closely to that of the isostructural superconductor CeCoIn₅ whose 4*f* electron from cerium contributes to the Fermi surface volume¹⁶. A localized to delocalized transition in the 4*f*-electron configuration is expected in a model of criticality in which extended and localized fluctuations

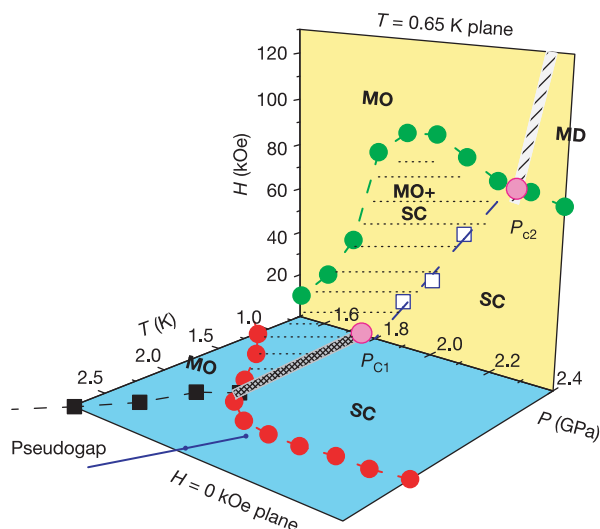


Figure 4 | The field-temperature-pressure phase diagram of the heavy fermion superconductor CeRhIn₅. The temperature–pressure plane is at $H = 0$ kOe and the H – P plane is for $T = 650$ mK. In the temperature–pressure plane, the magnetically ordered (MO) state of Ce $4f$ moments is preferred at low pressure. With increasing pressure, an SC phase appears and coexists with the MO phase when $P < 1.77$ GPa. For $1.77 < P < 2.3$ GPa, an SC phase is only found in zero field, but applied magnetic field induces a MO phase in the SC state. The blue line is a proposed pseudogap line¹³. In the H – P plane, upper critical fields H_{c2} , where superconductivity is totally depressed owing to the overlap of magnetic vortex cores, are represented by green circles. Quantum phase transitions between the pure SC phase and the coexistence phase of H -induced magnetism and SC are shown as open squares. The hatched grey line delineates a boundary between the MO phase and a magnetically disordered (MD) phase in the normal state. This boundary is defined by field-dependent specific heat and de Haas–van Alphen measurements, which were made at milliKelvin temperatures and fields $88 < H < 169$ kOe (ref. 16). P_{c1} is a quantum phase transition point between SC + MO and SC phases at zero magnetic field. P_{c2} is a tetracritical point where the H_{c2} line and the MO to MD lines cross. Experimental data constrain P_{c2} to within ± 0.05 GPa. The dashed line between P_{c1} and P_{c2} is a fit to the data as described in the text.

coexist at a quantum critical point²⁰. This model, however, does not include the role of superconductivity.

Neutron-diffraction experiments also have revealed field-induced magnetic order in the superconducting state of the high- T_c compound La_{1.9}Sr_{0.1}CuO₄ (ref. 21). Motivated by these observations, Demler and co-workers proposed a model that assumes the superconductor is near a quantum-phase transition to a state with microscopic coexistence of superconducting and magnetic orders^{6,7}. When the magnetic field penetrates an unconventional superconductor in which the superconducting energy gap has nodes on the Fermi surface, field-induced quantized vortices have an AFM ground state that suppresses superconductivity around the vortices. The suppression of the superconducting order enhances the competing AFM order even outside of the normal vortex cores, thus delocalizing magnetic correlations and creating microscopic coexistence of the AFM and superconducting orders. Repulsive coupling between AFM and superconducting (SC) orders that can be tuned either by chemical substitution or pressure tips the balance between the two competing ground states, leading to a quantum phase transition among the pure AFM phase, the AFM + SC coexisting phase, and the pure SC phase. This model accounts for the evolution of magnetic order in La_{1.9}Sr_{0.1}CuO₄ and its strengthening with increasing field²¹.

This model^{6,7} further predicts a line of quantum phase transitions between the AFM + SC and SC phases as a function of a control parameter δ . Taking pressure as the control parameter, this model

predicts: $H/H_{c2}^0 \approx 1 - \gamma[1 - \alpha(P - P_{c1})]$ for $H/H_{c2}^0 > 0.1$, where H_{c2}^0 is the upper critical field at a tetracritical point (P_{c2} in Fig. 4), γ is a numerical constant, and α is a proportionality between pressure and a repulsive coupling constant ($\delta = \alpha P$). A least-squares fit of this relationship to the open squares in Fig. 4 gives $\alpha = 1.99$, $\gamma = 1.11$, and $P_{c1} = 1.75$ GPa (dashed line in the field–pressure plane of Fig. 4). The numerical constant γ is in good agreement with that ($\gamma = 1.2$) obtained from a numerical solution of this model⁷. We also obtain P_{c2} ($H = H_{c2}^0$) = 2.25 GPa, which is very close to the pressure at which the effective mass of charge carriers diverges¹⁶. The H -linear proportionality of the magnetic entropy (Fig. 3a inset) and the microscopic coexistence of AFM and superconductivity in CeRhIn₅ are consistent with this model^{6,7}, which considers AFM order as the competing ground state of superconductivity. A possible explanation for the above phenomenological description of CeRhIn₅ is that the presence of superconductivity strongly inhibits a mechanism by which spins communicate, such as the Rudermann–Kittel–Kasuya–Yoshida (RRKY) interaction, which then may explain why magnetism is hidden in zero field by superconductivity when $T_c > T_N$.

Similarities between the high- T_c copper oxides^{21–25} and CeRhIn₅ suggest that some phenomena they exhibit may be ubiquitous features of magnetically mediated superconductivity. The model of criticality that accounts for our data is not specific to the microscopic origin of unconventional superconductivity or of quantum criticality. Antiferromagnetism is due to localized $4f$ -electrons and quantum criticality is associated with a localized-to-delocalized transition in the $4f$ -configuration of CeRhIn₅, but this is not an appropriate description of copper oxide physics, nor possibly of all heavy-electron compounds. Consequently, within this model, the mechanism of unconventional superconductivity may differ in detail from system to system, even though magnetism is a common denominator.

Received 10 November 2005; accepted 2 January 2006.

- Muller, K.-H. & Narozhnyi, V. N. Interaction of superconductivity and magnetism in borocarbide superconductors. *Rep. Prog. Phys.* **64**, 943–1008 (2001).
- Moriya, T. & Ueda, K. Antiferromagnetic spin fluctuations and superconductivity. *Rep. Prog. Phys.* **66**, 1299–1341 (2003).
- Curro, N. J. et al. Unconventional superconductivity in PuCoGa₅. *Nature* **434**, 622–625 (2005).
- Coleman, P. & Schofield, A. J. Quantum criticality. *Nature* **433**, 226–229 (2005).
- Mathur, N. D. et al. Magnetically mediated superconductivity in heavy fermion compounds. *Nature* **394**, 39–43 (1998).
- Demler, E., Sachdev, S. & Zhang, Y. Spin-ordered quantum transitions of superconductors in a magnetic field. *Phys. Rev. Lett.* **87**, 067202 (2001).
- Zhang, Y., Demler, E. & Sachdev, S. Competing orders in magnetic field: spin and charge order in the cuprate superconductors. *Phys. Rev. B* **66**, 094501 (2002).
- Fujita, M. et al. Magnetic and superconducting phase diagram of electron-doped Pr_{1-x}La_xCe_{1-x}CuO₄. *Phys. Rev. B* **67**, 014514 (2003).
- Demler, E., Hanke, W. & Zhang, S.-C. SO(5) theory of antiferromagnetism and superconductivity. *Rev. Mod. Phys.* **76**, 909–974 (2004).
- Lefebvre, S. et al. Mott transition, antiferromagnetism, and unconventional superconductivity in layered organic superconductors. *Phys. Rev. Lett.* **85**, 5420–5423 (2000).
- Hegger, H. et al. Pressure-induced superconductivity in quasi-2D CeRhIn₅. *Phys. Rev. Lett.* **84**, 4986–4989 (2000).
- Nicklas, M. et al. Two superconducting phases in CeRh_{1-x}Ir_xIn₅. *Phys. Rev. B* **70**, 020505R (2004).
- Mito, T. et al. Coexistence of antiferromagnetism and superconductivity near the quantum criticality of the heavy-fermion compound CeRhIn₅. *Phys. Rev. Lett.* **90**, 077004 (2003).
- Fisher, R. A. et al. Specific heat of CeRhIn₅: pressure-driven evolution of the ground state from antiferromagnetism to superconductivity. *Phys. Rev. B* **65**, 224509 (2002).
- Llobet, A. et al. Magnetic structure of CeRhIn₅ as a function of pressure and temperature. *Phys. Rev. B* **69**, 024403 (2004).
- Shishido, H. et al. A drastic change of the Fermi surface at a critical pressure in CeRhIn₅: dHvA study under pressure. *J. Phys. Soc. Jpn* **74**, 1103–1106 (2005).
- Bao, W. et al. Incommensurate magnetic structure of CeRhIn₅. *Phys. Rev. B* **62**, R14621 (2000).

18. Cornelius, A. L. *et al.* Field-induced magnetic transitions in the quasi-two-dimensional heavy-fermion antiferromagnets $\text{Ce}_n\text{RhIn}_{3n+2}$ ($n = 1$ or 2). *Phys. Rev. B* **64**, 144411 (2001).
19. Knebel, G. *et al.* High-pressure phase diagrams of CeRhIn_5 and CeCoIn_5 studied by ac calorimetry. *J. Phys. Condens. Matter* **16**, 8905–8922 (2004).
20. Si, Q. *et al.* Locally critical quantum phase transitions in strongly correlated metals. *Nature* **413**, 804–808 (2001).
21. Lake, B. *et al.* Antiferromagnetic order induced by an applied magnetic field in a high-temperature superconductor. *Nature* **415**, 299–302 (2002).
22. Lake, B. *et al.* Spins in the vortices of a high-temperature superconductor. *Science* **291**, 1759–1762 (2001).
23. Khaykovich, B. *et al.* Field-induced transition between magnetically disordered and ordered phases in underdoped $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. B* **71**, 220508R (2005).
24. Kang, H. J. *et al.* Antiferromagnetic order as the competing ground state in electron-doped $\text{Nd}_{1.85}\text{Ce}_{0.15}\text{CuO}_4$. *Nature* **423**, 522–525 (2003).
25. Panagopoulos, C. *et al.* Evidence for a generic quantum transition in high- T_c cuprates. *Phys. Rev. B* **66**, 064501 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank Y. K. Bang, A. V. Balatsky and N. J. Curro for discussions. Work at Los Alamos National Laboratory was performed under the auspices of the United States Department of Energy Office of Science. H.Q.Y. acknowledges an ICAM postdoctoral fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.P. (tuson@lanl.gov).

The influence of the surface migration of gold on the growth of silicon nanowires

J. B. Hannon¹, S. Kodambaka¹, F. M. Ross¹ & R. M. Tromp¹

Interest in nanowires continues to grow, fuelled in part by applications in nanotechnology^{1–5}. The ability to engineer nanowire properties makes them especially promising in nanoelectronics^{6–9}. Most silicon nanowires are grown using the vapour-liquid-solid (VLS) mechanism, in which the nanowire grows from a gold/silicon catalyst droplet during silicon chemical vapour deposition^{10–13}. Despite over 40 years of study, many aspects of VLS growth are not well understood. For example, in the conventional picture the catalyst droplet does not change during growth, and the nanowire sidewalls consist of clean silicon facets^{10–13}. Here we demonstrate that these assumptions are false for silicon nanowires grown on Si(111) under conditions where all of the experimental parameters (surface structure, gas cleanliness, and background contaminants) are carefully controlled. We show that gold diffusion during growth determines the length, shape, and sidewall properties of the nanowires. Gold from the catalyst droplets wets the nanowire sidewalls, eventually consuming the droplets and terminating VLS growth. Gold diffusion from the smaller droplets to the larger ones (Ostwald ripening) leads to nanowire diameters that change during growth. These results show that the silicon nanowire growth is fundamentally limited by gold diffusion: smooth, arbitrarily long nanowires cannot be grown without eliminating gold migration.

Our experiments were carried out in two separate ultrahigh vacuum electron microscopes: the IBM Low-Energy Electron Microscope (LEEM)¹⁴ and a Hitachi UHV H-9000 Transmission Electron Microscope (UHVTEM). The nanowires were grown on Si(111) cleaned in ultrahigh vacuum by repeated flashing to 1,200 °C. In the LEEM experiments two monolayers of Au were deposited on the substrates (at room temperature) from a resistively heated Knudsen cell. Upon heating to 600 °C, the Au layer partially dewetted, forming Au/Si droplets on a smooth Au wetting layer. Selected-area electron diffraction measurements (in the LEEM) indicated that the surface

between the droplets had the $\beta - (\sqrt{3} \times \sqrt{3})$ structure corresponding to 1 monolayer of Au on Si(111)^{15,16}. *In situ* imaging with LEEM showed that Au at the surface is highly mobile at 600 °C, with large catalyst droplets growing at the expense of smaller ones (Fig. 1).

After forming the Au/Si droplets, we grew Si nanowires by exposing the surface (at 600 °C in the LEEM) to a mixture of 20% disilane and 80% He at a pressure of 5×10^{-5} torr. Following 30 minutes of growth, the surface was imaged (at 600 °C) using LEEM. Diffraction measurements indicated a well-ordered $\beta - (\sqrt{3} \times \sqrt{3})$ structure between the nanowires, consistent with 1 monolayer of Au on Si(111)¹⁵. SEM images recorded after the disilane exposure are shown in Fig. 2a and b. Au/Si catalyst droplets are visible at the ends of all of the longer nanowires. These nanowires have a uniform length of about 750 nm. A small fraction of the nanowires (indicated by arrows in Fig. 2a) are cone-shaped, shorter, and have no catalyst

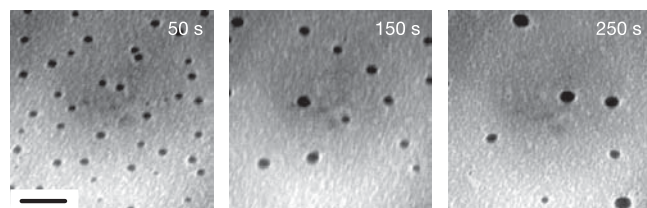


Figure 1 | 5-eV LEEM images of the coarsening of Au catalyst droplets on Si(111). Diffraction measurements indicate that the surface in between the droplets is covered by approximately one monolayer of Au. Each image is labelled by the total annealing time at 600 °C. The number of droplets decreases with time as larger droplets grow at the expense of smaller ones. Scale bar for all panels, 1 μm .

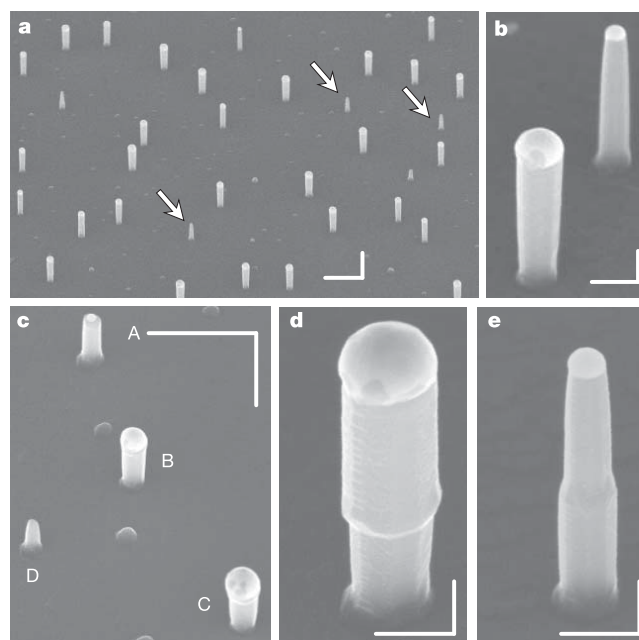


Figure 2 | Scanning electron microscope images of Si nanowires grown on Si(111) recorded with a 42° angle of incidence. **a**, Shorter, cone-shaped nanowires without visible catalyst droplets are indicated by arrows. **b**, Coarsening of the catalyst particles leads to wires with diameters that either grow or shrink during nanowire growth. **c**, Nanowires imaged after 35 min of annealing at 600 °C following nanowire growth for 35 min at 600 °C. **d, e**, Nanowires imaged after 30 min of growth at 600 °C, followed by annealing at 650 °C for 20 min, and a final growth period of 25 min at 600 °C. Scale bars: **a, c**, 1 μm ; **b, d, e**, 200 nm.

¹IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598, USA.

droplets. The results are qualitatively similar to those of Givargizov¹², who observed that shorter nanowires often had no catalyst droplets. High-resolution imaging reveals that the sidewalls of the nanowires are not perfectly straight. As shown in Fig. 2b, the diameters of some wires decreased with time during growth, while those of others increased. These results suggest that the sizes of the catalyst droplets change during nanowire growth at 600 °C. Given that the Au/Si droplets on the surface coarsen at 600 °C before growth, it is likely that the droplets continue to coarsen during nanowire growth. That is, during nanowire growth Au preferentially migrates from the smaller catalyst droplets to the larger ones to reduce the total droplet surface area. In this process some of the droplets coarsen away completely, leading to the shorter, cone-shaped nanowires shown in Fig. 2a.

This hypothesis can be tested by continuing to anneal the sample at 600 °C after nanowire growth. The results for a sample annealed at 600 °C for 35 min (following nanowire growth for 35 minutes also at 600 °C) are shown in Fig. 2c. The wires labelled A, B, and C all have the same length. They also have approximately the same diameter, indicating that they grew from catalyst droplets of similar size. However, after the post-growth annealing, the diameters are very different. The droplet on wire C has clearly grown, while that on wire A has shrunk. The droplet on wire B is essentially unchanged. The nanowire labelled D is shorter than the others, suggesting that its droplet coarsened away during nanowire growth. The coarsening can be exploited to introduce a discontinuous change in the nanowire diameters. For example, a second stage of growth can be carried out on a post-annealed sample (like that shown in Fig. 2c). During the post-growth anneal the catalyst droplets coarsen and change size, so that when growth is resumed, the nanowires grow with either a larger or smaller diameter. Examples of both types of nanowires are shown in Fig. 2d and e.

The coarsening of the catalyst droplets during nanowire growth

can be observed directly in the UHVTEM. In these experiments the sample was mounted in a reflection electron microscopy geometry¹⁷. The Si(111) surface was cleaned by flashing *in situ* to 1,200 °C and about two monolayers of Au were then evaporated at room temperature. The sample was heated to 550 °C or above to form droplets, and nanowires were then grown in pure disilane at 10^{-6} torr and temperatures between 550 and 675 °C. As the nanowires grew away from the surface they could be imaged in transmission mode. A sequence of images from a growth experiment at 665 °C is shown in Fig. 3. During growth the droplet labelled A grows larger, while the one labelled B decays and eventually disappears.

These results clearly show that Au can diffuse from one catalyst droplet to another during growth. That is, Au migrates over the surface as well as up and down the nanowire sidewalls. Diffraction measurements in the LEEM show that the surface remains covered by one monolayer of Au during nanowire growth. Separate experiments performed on Si(001) substrates gave similar results. The diffusion of Au between droplets suggest that the nanowire sidewalls are also

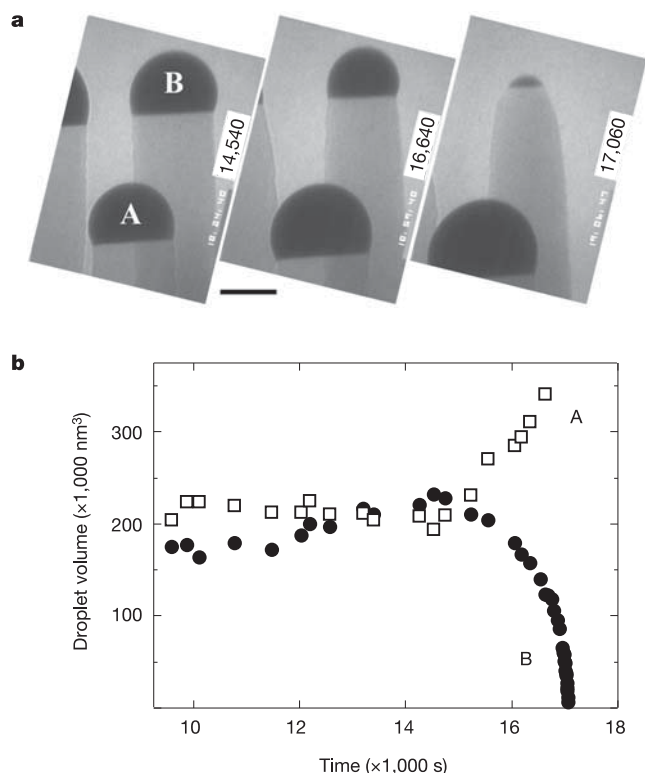


Figure 3 | *In situ* UHVTEM images recorded during the growth of Si nanowires at 655 °C in 10^{-6} torr disilane. a, Three images labelled by the time (in seconds) after the start of growth. Scale bar: 50 nm. **b**, Measured volume versus time for the two catalyst droplets labelled in **a**.

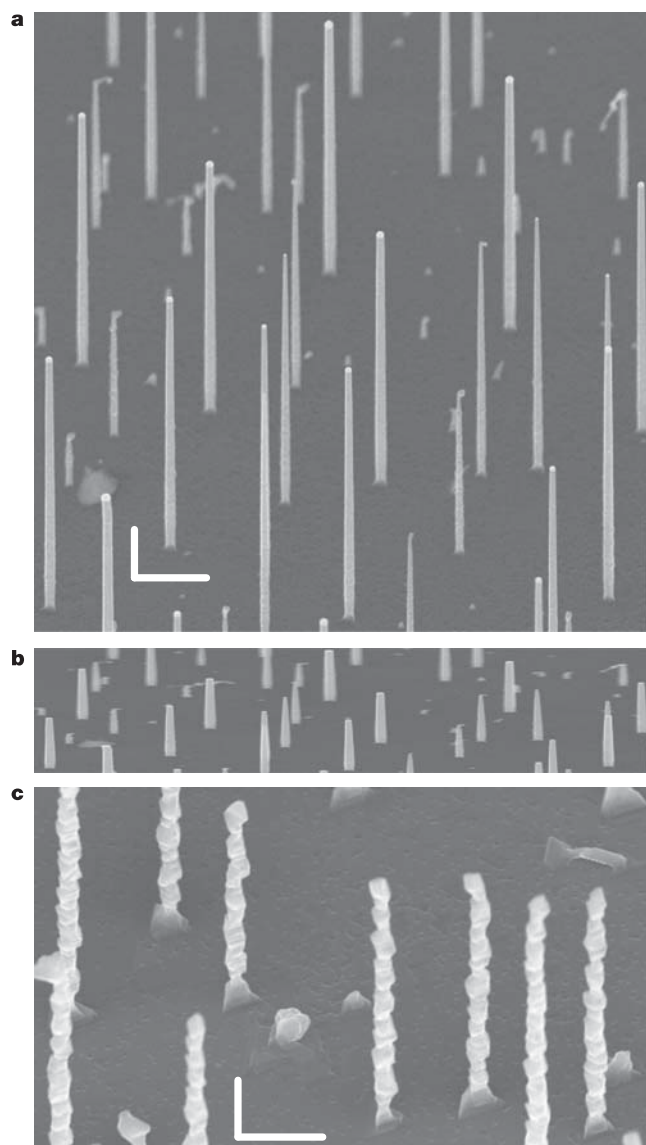


Figure 4 | SEM images of Si nanowires. a, Si nanowires grown for 1 h at 600 °C in a 20% disilane and 80% He solution at a pressure of 5×10^{-4} torr. **b**, Same image with vertical scaling reduced by a factor of five to highlight the sidewall tapering. **c**, Nanowires grown for 2 h at 600 °C at a pressure of 5×10^{-4} torr. Scale bars: 1 μm .

covered by a significant amount of Au. Wetting of the sidewalls fundamentally limits how long the nanowires can grow: extending the nanowire length requires a transfer of Au from the catalyst droplet to the sidewall. Au in the catalyst droplet will eventually be completely consumed and VLS growth will terminate.

We have investigated this possibility by growing longer nanowires in the LEEM at a higher disilane pressure. Specifically, we produced 100–200 nm Au/Si catalyst droplets on the $\sqrt{3} \times \sqrt{3}$ Au wetting layer by heating a two-monolayer Au film to 600 °C for several minutes. We then exposed the surface (at 600 °C) to a mixture of 20% disilane in He at a total pressure of 5×10^{-4} torr for 1 h. Scanning electron microscope images after growth are shown in Fig. 4a. Most of the nanowires are 5–6 μm in length and most still have a droplet at the end. However, what is notable is that all of the catalyst droplets have decreased in size during growth, resulting in a gradual tapering of all of the nanowires. The tapering is more easily seen in Fig. 4b, where the vertical scale has been reduced to highlight changes in the nanowire diameters. Clearly, there is a net transfer of Au out of the catalyst droplets during growth. The probable reason for this is that Au wets the nanowire sidewalls. Given the fact that Au wets both Si(001) and Si(111), migration of Au from the droplets to the sidewalls is perhaps expected. Furthermore, the coarsening of the droplets during growth is easier to understand if both the substrate and sidewalls are covered by Au. From our experiments we cannot exclude the possibility that Au becomes incorporated into the interior of the nanowire. However, given the extremely low solubility of Au in bulk Si (fewer than 10^{-9} Au atoms per Si atom at 600 °C; ref. 18), significant incorporation is unexpected. If no bulk incorporation is assumed, the amount of Au coating the sidewalls can be estimated from the length of the nanowire. If Au from the (hemispherical) catalyst droplet is transferred to the sidewall, the droplet will be completely consumed when the nanowire length, L , is equal to:

$$L = r^2 / 2a\theta \quad (1)$$

where r is the radius of the nanowire at the base, θ is the average coverage of Au on the sidewall, and a^3 is the atomic volume of Au ($a \approx 0.26$ nm). Some of the long nanowires in Fig. 4 have no catalyst droplet, presumably because their droplets were consumed just before growth ended. These wires can be used to estimate θ . Typical values of $L = 5$ – $6 \mu\text{m}$ and $r = 50$ – 70 nm suggest that θ is in the range of 1.0 to 1.5 monolayers, which is reasonable given the approximate nature of the calculation. We note that this analysis implies that very small nanowires ($r < 10$ nm) cannot grow longer than a few 100 nm before their catalyst droplets are consumed.

During VLS growth, when catalyst droplets are present, the nanowire sidewalls are relatively smooth (Fig. 4a). This suggests that the rate of VLS growth is fast compared to the rate of chemical vapour deposition (CVD) growth on the Au-covered sidewall. However, once the droplets are consumed, VLS growth terminates while direct CVD growth of Si on the sidewalls continues, dramatically altering the sidewall morphology, as shown in Fig. 4c. In this experiment nanowires were grown for 2 h at 600 °C in a total pressure of 5×10^{-4} torr. After growth, no catalyst droplets are found on any of the nanowires. From the length of the nanowires, we estimate that the droplets were consumed after the first hour of growth. Direct CVD growth of Si on the sidewalls is not smooth: the formation of facets along the length of the wire leads to very rough, undulating sidewall profiles.

We have shown that Au diffusion fundamentally limits the length, shape and sidewall properties of Si nanowires. While these observations provide a baseline for the understanding of VLS Si nanowire growth, we also acknowledge that very long, thin nanowires are

grown routinely in dedicated nanowire CVD systems. These nanowires have constant diameters and have no Au decorating the sidewalls. Clearly there are critical differences between low-pressure CVD in ultrahigh vacuum on clean substrates and conventional CVD at higher pressure that prevent Au migration in the latter case. One inescapable difference is cleanliness. It may be that adsorption of contaminants from the precursor gas reduces Au migration. Several investigations have shown that growth in the presence of oxygen, for example, strongly influences nanowire properties^{19,20}, and our own recent measurements confirm this for low-pressure CVD growth. However, regardless of what the exact differences between low-pressure and high-pressure CVD are, the results presented here show that the growth of fixed-diameter, arbitrarily long Si nanowires depends critically on limiting the surface migration of Au. A detailed and critical assessment of the VLS CVD process will be required to determine how the limits outlined in this paper can be circumvented to enable the routine, reproducible growth of technologically relevant Si nanowires.

Received 8 August; accepted 20 December 2005.

Published online 29 January 2006.

- Huang, Y. *et al.* Logic gates and computation from assembled nanowire building blocks. *Science* **294**, 1313–1317 (2001).
- Duan, X., Huang, Y., Cui, Y., Wang, J. & Lieber, C. M. Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices. *Nature* **409**, 66–69 (2001).
- Duan, X., Huang, Y. & Lieber, C. M. Nonvolatile memory and programmable logic from molecule-gated nanowires. *Nano Lett.* **2**, 487–490 (2002).
- Cui, Y., Zhong, Z., Wang, D., Wang, W. U. & Lieber, C. M. High performance silicon nanowire field effect transistors. *Nano Lett.* **3**, 149–152 (2003).
- Wu, Y. *et al.* Controlled growth and structure of molecular-scale silicon nanowires. *Nano Lett.* **4**, 433–436 (2004).
- Lauhon, L. J., Gudiksen, M. S., Wang, D. & Lieber, C. M. Epitaxial core-shell and core-multishell nanowire heterostructures. *Nature* **420**, 57–61 (2002).
- Björk, M. *et al.* One-dimensional heterostructures in semiconductor nanowhiskers. *Appl. Phys. Lett.* **80**, 1058–1060 (2002).
- Björk, M. *et al.* One-dimensional steepenclase for electrons realized. *Nano Lett.* **2**, 87–89 (2002).
- Gudiksen, M. S., Lauhon, L. J., Wang, J., Smith, D. C. & Lieber, C. M. Growth of nanowire superlattice structures for nanoscale photonics and electronics. *Nature* **415**, 617–620 (2002).
- Wagner, R. S. & Ellis, W. C. Vapor-liquid-solid mechanism of single crystal growth. *Appl. Phys. Lett.* **4**, 89–90 (1964).
- Wagner, R. S. in *Whisker Technology* (ed. Levitt, A. P.) 47–119 (Wiley-Interscience, New York, 1970).
- Givargizov, E. I. Fundamental aspects of VLS growth. *J. Cryst. Growth* **31**, 20–30 (1975).
- Law, M., Goldberger, J. & Yang, P. Semiconductor nanowires and nanotubes. *Annu. Rev. Mater. Res.* **34**, 83–122 (2004).
- Tromp, R. M. & Reuter, M. C. Design of a new photoemission/low-energy electron microscope for surface studies. *Ultramicroscopy* **36**, 99–106 (1991).
- Nagao, T. *et al.* Structural phase transitions of Si(111) – $\sqrt{3} \times \sqrt{3}$ – R30°-Au: Phase transitions in domain-wall configurations. *Phys. Rev. B* **57**, 10100–10109 (1998).
- Zhang, H. M., Balasubramanian, T. & Uhrberg, R. I. G. Core-level photoelectron spectroscopy study of the Au/Si(111) 5×2 , $\alpha - \sqrt{3} \times \sqrt{3}$, $\beta - \sqrt{3} \times \sqrt{3}$, and 6×6 surfaces. *Phys. Rev. B* **65**, 035314 (2001).
- Yagi, K. Reflection electron microscopy—studies of surface structures and surface dynamic processes. *Surf. Sci. Rep.* **17**, 305–362 (1993).
- Stolwijk, N. A., Schuster, B., Hölzl, J., Mehrer, H. & Frank, W. Diffusion and solubility of gold in silicon. *Physica B* **116**, 335–342 (1983).
- Lee, S. T., Wang, N. & Lee, C. S. Semiconductor nanowires: synthesis, structure and properties. *Mater. Sci. Eng. A* **286**, 16–23 (2000).
- Kolb, F. M. *et al.* Analysis of silicon nanowires grown by combining SiO evaporation with the VLS mechanism. *J. Electrochem. Soc.* **151**, G472–C475 (2004).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.B.H. (jbhannon@us.ibm.com).

LETTERS

Uplift, thermal unrest and magma intrusion at Yellowstone caldera

Charles W. Wicks¹, Wayne Thatcher¹, Daniel Dzurisin² & Jerry Svarc¹

The Yellowstone caldera, in the western United States, formed ~640,000 years ago when an explosive eruption ejected ~1,000 km³ of material¹. It is the youngest of a series of large calderas that formed during sequential cataclysmic eruptions that began ~16 million years ago in eastern Oregon and northern Nevada. The Yellowstone caldera was largely buried by rhyolite lava flows during eruptions that occurred from ~150,000 to ~70,000 years ago¹. Since the last eruption, Yellowstone has remained restless, with high seismicity, continuing uplift/subsidence episodes with movements of ~70 cm historically² to several metres since the Pleistocene epoch³, and intense hydrothermal activity. Here we present observations of a new mode of surface deformation in Yellowstone, based on radar interferometry observations from the European Space Agency ERS-2 satellite. We infer that the observed pattern of uplift and subsidence results from variations in the movement of molten basalt into and out of the Yellowstone volcanic system.

In a previous satellite interferometric synthetic aperture radar (InSAR) study of Yellowstone⁴, interferograms from 1992 to 1997 revealed a change from caldera-wide subsidence that began in 1985 (ref. 5) to uplift that began in 1995 and by 1997 involved the whole caldera floor (Fig. 1) as well as the area of uplift shown in Fig. 2a. Because the area of uplift in Fig. 2a is under the north caldera rim, we refer to it as the NUA (north rim uplift anomaly). Surface movements from 1996 to 2002 have proven to be even more dynamic (Fig. 2), with important implications for the nature of the magmatic plumbing of Yellowstone caldera and large caldera systems in general. Campaign mode GPS measurements from 1995 to 2000 (ref. 6) first revealed that NUA had become an isolated area of uplift, in agreement with InSAR observations (Fig. 2) that are drawn from all available satellite radar data. As NUA continued to rise after 1995, vertical motion of the caldera floor connecting the two resurgent domes (Sour Creek, SC, and Mallard Lake, ML, in Fig. 1) changed from uplift to subsidence between late 1997 and early 1998. NUA continued to inflate as the caldera floor subsided until 2002 at which time both movements ceased, or at least paused (Fig. 2b–d).

In order to model a deformation source for the entire inflation episode at NUA, we formed the interferogram in Fig. 3a by summing the interferograms in Fig. 2a–c. The total amount of volume added by the modelled inflating sill beneath NUA (see Supplementary Information) shown in Fig. 3 is 0.06–0.1 km³. The best-fit model synthetic interferogram is shown in Fig. 3b and the residual is shown in Fig. 3c. In Fig. 3d we show a profile through Y–Y' (Fig. 3a) that passes through the peak uplift at NUA and the peak subsidence at SC. By all appearances, the uplift at NUA and the subsidence of the caldera floor are linked. Therefore, any model that explains the uplift should also explain the subsidence.

Past episodes of uplift and subsidence in the caldera have been

attributed to various combinations of the following two processes taking place beneath the caldera. (1) Pressurization and de-pressurization of an alternately self-sealed and leaking hydrothermal fluid reservoir that traps volatiles exsolved from a crystallizing rhyolitic magma⁷. (2) Movement, formation and crystallization of rhyolitic and/or basaltic magma.

Ingebritsen *et al.*⁸ noted that chloride flux measurements at Yellowstone—which provide a measure of the hydrothermal heat

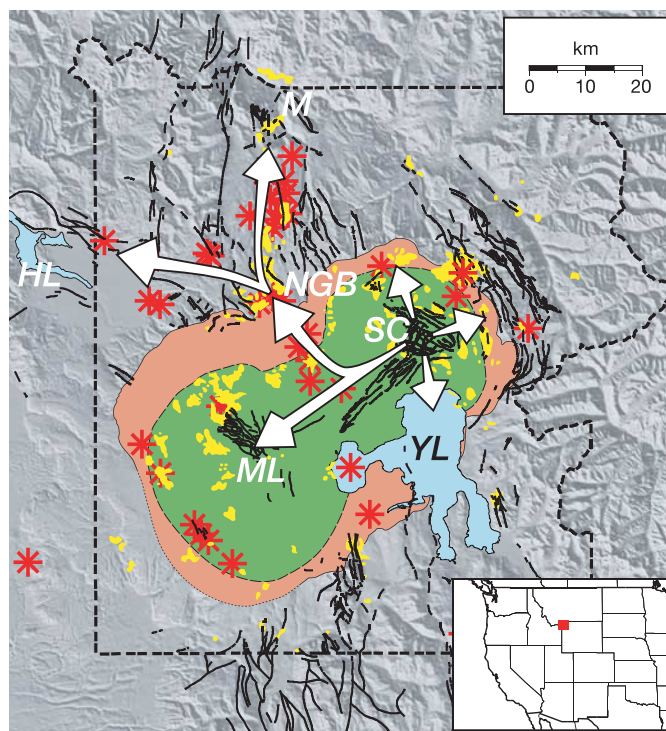


Figure 1 | A map of structural, thermal and volcanic features in and around Yellowstone caldera. (Map taken from Christiansen¹.) The red symbols mark volcanic centres that erupted after the caldera-forming event 640 kyr ago. The areas of known past or present thermal activity are coloured yellow. The ring-fracture zone of the caldera is shown green, and the slumped zone between the ring-fracture zone and the best estimate of the caldera rim is shown salmon. The park boundary is the dashed black line. Faults active in the Quaternary are marked with black lines. The labelled features are Norris Geyser basin (NGB), Mammoth Hot Springs (M), Sour Creek dome (SC), Mallard Lake dome (ML), Hebgen Lake (HL) and Yellowstone Lake (YL). The white arrows show interpreted magma migration paths. The red square in the inset map (bottom right) shows the location of the study area.

¹US Geological Survey, MS 977, Menlo Park, California 94555, USA. ²US Geological Survey, David A. Johnston Cascades Volcano Observatory, 1300 SE Cardinal Court, Bldg 10, Suite 100, Vancouver, Washington 98683, USA.

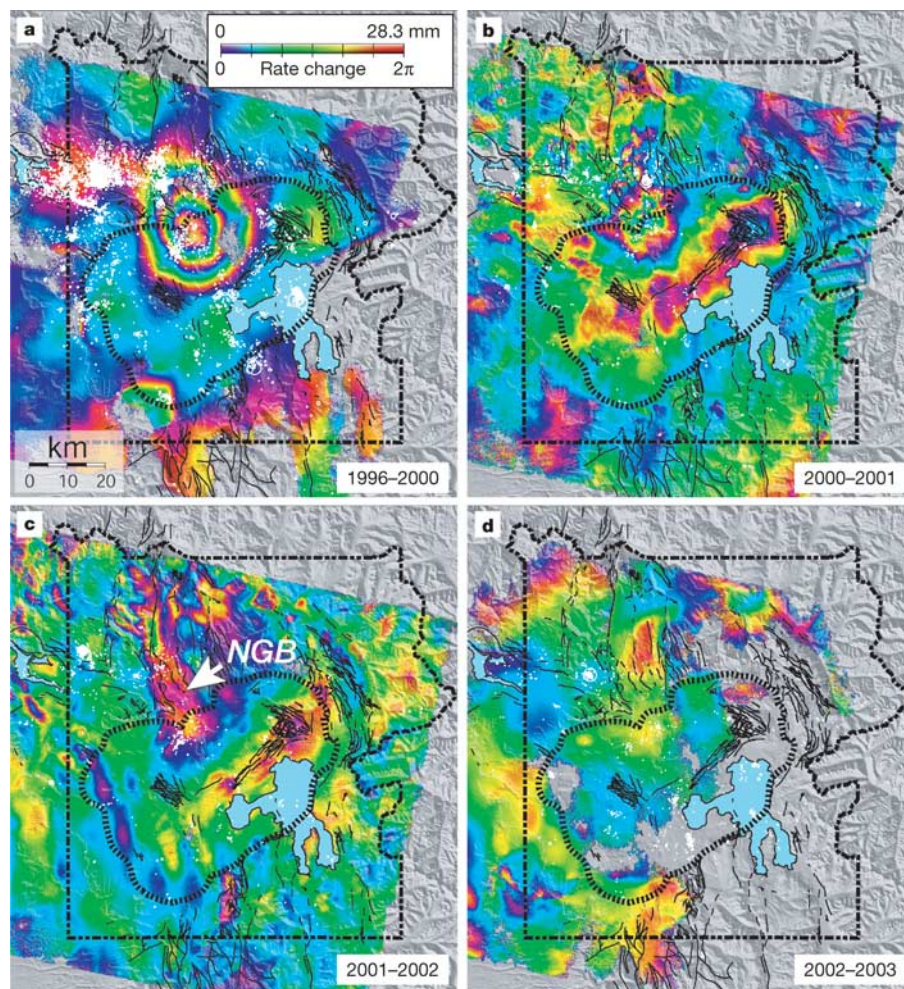


Figure 2 | Four interferograms showing the deformation during the episode of uplift at NUA. A colour change from violet to blue to green to yellow to red marks an increase in the range (distance from the satellite to points on the ground) of 28.3 mm. The white circles represent epicentres of earthquakes recorded during the time interval spanned by each interferogram. The interferograms have been generated using European Space Agency ERS-2 data (see Supplementary Information) and the two-pass method of interferometry²⁴. The extensive double dash length broken line in each panel shows the boundary of Yellowstone National Park. The short dash length broken line in each panel (within the park boundary) shows the approximate location of the 640,000-year-old caldera rim. **a**, Summer 1996 to summer 2000 interferogram. Although the caldera floor appears to have subsided only slightly, this period includes about 30 mm of caldera-wide uplift from 1996 to 1997 (ref. 4). Therefore, more than 30 mm of subsidence of the caldera floor occurred between the ML and SC resurgent domes (Fig. 1) from 1997 to 2000. **b**, Summer 2000 to summer 2001 interferogram. **c**, Summer 2001 to summer 2002 interferogram. The arrow labelled NGB marks the location of Norris Geyser basin. **d**, Summer 2002 to summer 2003 interferogram.

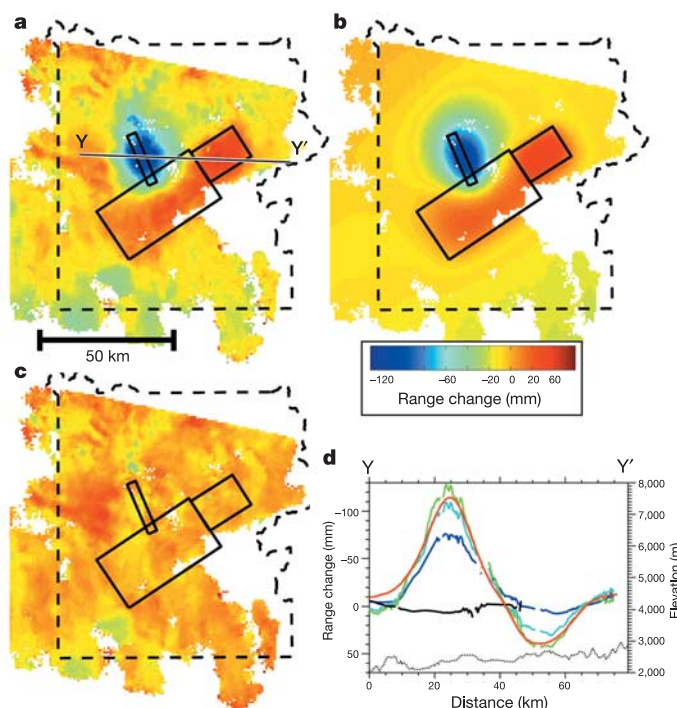


Figure 3 | Observed and modelled uplift at NUA and subsidence of the caldera floor. The black outlines are the surface projections of a north-northwest-trending expanding sill, and two northeast-trending contracting sills. The dashed line is the outline of Yellowstone National Park. **a**, A stacked interferogram formed by summing unwrapped versions of the interferograms in Fig. 2a–c. **b**, Synthetic interferogram from best-fit model. **c**, Residual interferogram formed by subtracting the synthetic interferogram (**b**) from the observed interferogram (**a**). **d**, Deformation profiles from Y to Y'. The colours denote: blue, 1996–2000 deformation; cyan, 1996–2001 deformation; green, 1996–2002 deformation; black, 2002–2003 deformation; and red, deformation from best-fit model (**b**). Elevation along the profile is shown by the black dotted line.

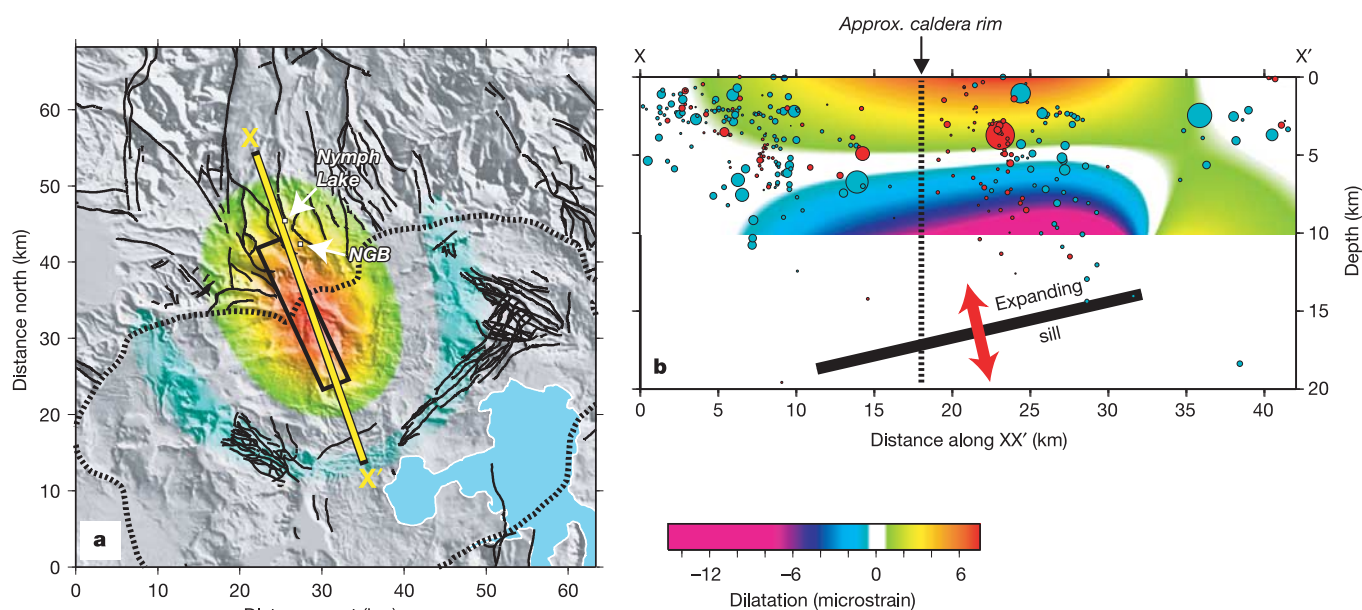


Figure 4 | Dilatation calculated from the inflating sill in Fig. 3. (Calculations were performed using Coulomb 2.5; ref. 25.) **a**, Dilatation at the surface resulting from the NUA uplift episode. The arrow labelled 'NGB' marks the location of Norris Geyser basin, and the arrow labelled 'Nymph Lake' marks the location of the newly formed line of fumaroles near Nymph Lake. The black lines are mapped faults active in the Quaternary¹. The black rectangle is the surface projection of the best-fit expanding sill. The broken line shows the approximate location of the 640,000-year-old caldera rim. **b**, Cross-section through X–X' in **a** resulting from the uplift episode. The peak dilatation is just under 7 microstrain at the surface. The cyan circles show

earthquakes greater than $M = 0.0$ that occurred before the uplift episode (1 January 1992 through to 31 December 1997) and the red circles show earthquakes greater than $M = 0.0$ that occurred during the uplift episode (1 January 1998 through to 13 August 2003). Earthquakes 10 km each side of the X–X' line are projected onto the cross-section. Size of circles is scaled to earthquake magnitude. Note that for better visualization, this scaling is different for the red and cyan circles; for two earthquakes with the same magnitude (one cyan, one red), the red circle plots at twice the diameter of the cyan circle.

loss—do not show any obvious temporal changes that might relate to changes in deformation style. This led them to suggest that deep magmatic processes were more likely causes of uplift/subsidence cycles at Yellowstone. Because chloride is one of the volatile species that is exsolved from rhyolitic magma during crystallization, any rupture of a sealed hydrothermal reservoir might be expected to eventually produce a corresponding increase in chloride flux at the surface. The lack of chloride flux anomalies that are correlated to deformation episodes does not rule out a hydrothermal deformation source, but it is more easily compatible with a magmatic source for deformation.

We propose that the observed patterns of uplift and subsidence result from variations in what may be nearly continuous movement of molten basalt in to and out of the Yellowstone volcanic system. Increases in the rate of basaltic magma flux into the caldera from beneath SC (Fig. 1) favour inflation of the caldera, whereas decreases favour subsidence. Increases in the rate of basaltic magma flux out of the caldera near NGB (Fig. 1) favour subsidence of the caldera, whereas decreases favour inflation. The main driving forces moving the basaltic magma into the system are the integrated buoyancy of the magma and the vertical gradient in normal stress. The combination of extensional stress (Yellowstone is at the northeastern corner of Basin-and-Range extension) and high heat flow present in Yellowstone is expected to favour emplacement of magma at rheological boundaries^{9,10}. If the flux of magma is too great, however, it would tend to continue its near vertical ascent¹¹.

We interpret the beginning of caldera uplift in 1995 as the introduction of a pulse of basaltic magma below SC from a source in the upper mantle. In the subsequent year, the uplift spread across the entire caldera⁴, including the area of NUA, as the magma spread horizontally at a rheological boundary. Previous studies have inferred the presence of a partially molten rhyolitic body^{1,12–14} that would form a rheological boundary with an accompanying sharp upward

decrease in density. As the basalt spreads beneath the caldera, it loses heat to the overlying rock. This heat keeps the geothermal system, and thus the surface hydrothermal features, active.

The basaltic magma escapes the caldera system at the three-way intersection of the northern caldera boundary, the west-northwest-striking seismic belt east of the Hebgen Lake fault zone, and the north-trending Norris–Mammoth corridor (Fig. 1). The seismic belt is a rift-like zone of north–south extension^{15,16} that is a site of minor post-caldera volcanism and extends west to the site of the 1959 M_s (surface wave magnitude) 7.5 Hebgen Lake earthquake. The Norris–Mammoth corridor is a zone of recurrent normal faulting, post-caldera volcanism, and active thermal features that extends north to beyond Mammoth Hot Springs¹.

Magma accumulated beneath the north caldera boundary, leading to continued uplift at NUA even as the larger part of the caldera floor subsided, because, in our interpretation, the outlet was unable to fully accommodate the increased flux of basaltic magma. The flux of magma out of the Yellowstone system is controlled by extra-caldera tectonic activity acting on fractured rock bordering the northern caldera boundary. Tectonic strain can either enhance or restrict the flow of magma out of the caldera. The two largest earthquake swarms recorded in Yellowstone each accompanied, or slightly preceded, the change from caldera-wide uplift to subsidence in 1985, and the change from caldera-wide subsidence to a brief episode of caldera-wide uplift in 1995 (refs 4, 17). A similar scenario has been suggested at Loihi volcano in Hawaii¹⁸, where a 1996 earthquake swarm was associated with magma chamber drainage, and a 2001 earthquake swarm was associated with magma chamber filling. The shallower southeastern end of the dipping sill that models the inflation at NUA is at the same depth as the two deflating sills it intersects beneath the caldera floor. The dipping sill deepens to the north-northwest by ~ 7 km at the proposed outlet, effectively forming a trap for the now negatively buoyant magma. Inclusion of GPS data⁶ in a joint

inversion of GPS and InSAR data (Supplementary Information) yields a similarly oriented prolate spheroid as an allowable model. Adoption of the spheroid does not change the proposed path of magma migration, but rather the mode of migration, to one more pipe-like in nature. As the magma leaves the caldera, it could become negatively buoyant by way of three separate processes: (1) cooling, (2) crystallization and (3) degassing of CO₂. The denser magma might then be emplaced in a large 10-km-thick sill inferred in ref. 19.

Inflation beneath NUA has led to extensive dilatation of the upper crust (Fig. 4). In faulted and fractured areas, such as geyser basins, this could lead to a dramatic increase in permeability. The 2000–2001 interferogram (Fig. 2b) shows several small (2–5 km) areas of inflation with ~30–50 mm of peak amplitude north of NUA in the highly faulted, thermally active Mammoth–Norris corridor, which may be atmospheric delay artefacts. However, drawing on the results of a study by Hanssen *et al.*²⁰ and ground-based radar (see Supplementary Information), it is more likely that these are areas of local inflation. These areas are also absent in the 1996–2000 interferogram (Fig. 2a), and they have broadened and extended northward in 2001–2002 (Fig. 2c) to occupy the entire Mammoth–Norris corridor. We suggest that the dilatation opened new or healed fractures or increased permeability in existing fractures, resulting in better communication between the shallow thermal systems and the deeper geothermal reservoir, thus forming the small-scale areas of inflation.

Thermal disturbances in Norris Geyser basin (Figs 1, 4) are near-annual events that have been related to yearly water table lows²¹. The thermal disturbances have recently become more pronounced (see Supplementary Information), perhaps in response to dilatation from NUA. Ingebritsen and Rojstaczer²² have demonstrated that the permeability of a geyser's fracture zone conduit may be an important factor in the eruption frequency of a geyser. Dilatation from NUA is thus a possible mechanism for increasing permeability in the geyser conduits, thereby increasing geyser eruption frequency. Husen *et al.*²³ noted an increase in geyser eruption frequency within hours after surface waves from the 2002 Denali earthquake produced ~0.5 microstrain of dynamic strain at Yellowstone. We calculate strains an order of magnitude greater (>6 microstrain) applied over a time interval four orders of magnitude longer (~3–4 yr). Inertial forces are insignificant for the dilatation we calculate, but there must be competition between (1) opening of cracks by hydro-fracturing and (2) healing of cracks through mineral precipitation and annealing that, beginning in 2000, led to the manifestation of shallow responses to the dilatation.

The episode of accentuated thermal unrest in the near-annual disturbances from 2000–2003 is not unique in the recorded history of Yellowstone National Park²¹. It is unique, however, that for this episode we have been able to use InSAR to track changes in the deformation field in the park during the unrest that suggest a cause-and-effect relationship. Indeed, past episodes of accentuated thermal unrest during near-annual disturbances in Norris Geyser basin may also have been caused by dilatation related to uplift from magma accumulation at depth, but in the absence of geodetic monitoring, earlier deformation episodes would have gone undetected.

Received 22 February; accepted 30 November 2005.

- Christiansen, R. L. The Quaternary and Pliocene Yellowstone plateau volcanic field of Wyoming, Idaho, and Montana. *US Geol. Surv. Prof. Pap.* **729-G**, 1–145 (2001).
- Pelton, J. R. & Smith, R. B. Recent crustal uplift in Yellowstone National Park. *Science* **206**, 1179–1182 (1979).
- Pierce, K. L., Cannon, K. P., Meyer, G. A., Trebesch, M. J. & Watts, R. D. Post-glacial inflation–deflation cycles, tilting, and faulting in the Yellowstone Caldera based on Yellowstone lake shorelines. *US Geol. Surv. Open-File Rep.* **02–0142** 1–62 (2002).

- Wicks, C., Thatcher, W. & Dzurisin, D. Migration of fluids beneath Yellowstone caldera inferred from satellite radar interferometry. *Science* **282**, 458–462 (1998).
- Dzurisin, D., Savage, J. C. & Fournier, R. O. Recent crustal subsidence at Yellowstone caldera, Wyoming. *Bull. Volcanol.* **52**, 247–270 (1990).
- Meertens, C. M., Smith, R. B. & Puskas, C. M. Crustal deformation of the Yellowstone caldera from campaign and continuous GPS surveys, 1987–2000. *Eos* **81**, V22F–19 (2000).
- Fournier, R. O. Hydrothermal processes related to movement of fluid from plastic into brittle rock in the magmatic–epithermal environment. *Econ. Geol.* **94**, 1193–1211 (1999).
- Ingebritsen, S. E. *et al.* Time-variation of hydrothermal discharge at selected sites in the western United States: implications for monitoring. *J. Volcanol. Geotherm. Res.* **111**, 1–23 (2001).
- Rubin, A. M. Propagation of magma-filled cracks. *Annu. Rev. Earth Planet. Sci.* **23**, 287–336 (1995).
- Watanabe, T., Koyaguchi, T. & Seno, T. Tectonic stress controls on ascent and emplacement of magmas. *J. Volcanol. Geotherm. Res.* **91**, 65–78 (1999).
- Dahm, T. Numerical simulations of the propagation path and the arrest of fluid-filled fractures in the Earth. *Geophys. J. Int.* **141**, 623–638 (2000).
- Eaton, G. P. *et al.* Magma beneath Yellowstone National Park. *Science* **188**, 787–796 (1975).
- Miller, D. S. & Smith, R. B. *P* and *S* velocity structure of the Yellowstone volcanic field from local earthquake and controlled source tomography. *J. Geophys. Res.* **104**, 15105–15121 (1999).
- Husen, S., Smith, R. B. & Waite, G. P. Evidence for gas and magmatic sources beneath the Yellowstone volcanic field from seismic tomographic imaging. *J. Volcanol. Geotherm. Res.* **131**, 397–410 (2004).
- Savage, J. C., Lisowski, M., Prescott, W. H. & Pitt, A. M. Deformation from 1973 to 1987 in the epicentral area of the 1959 Hebgen Lake, Montana, earthquake ($M_s = 7.5$). *J. Geophys. Res.* **98**, 2145–2153 (1993).
- Waite, G. P. & Smith, R. B. Seismotectonics and stress field of the Yellowstone volcanic plateau from earthquake first-motions and other indicators. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002675 (2004).
- Waite, G. P. & Smith, R. B. Seismic evidence for fluid migration accompanying subsidence of the Yellowstone Caldera. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000586 (2002).
- Wolfe, C. J., Okubo, P. G., Ekström, G., Nettles, M. & Shearer, P. M. Characteristics of deep (≥ 13 km) Hawaiian earthquakes and Hawaiian earthquakes west of 155.55°W. *Geochem. Geophys. Geosyst.* **5**, doi:10.1029/2003GC000618 (2004).
- Humphreys, E. D., Dueker, K. G., Schutt, D. L. & Smith, R. B. Beneath Yellowstone: Evaluating plume and nonplume models using teleseismic images of the upper mantle. *GSA Today* **10**, 1–7 (2000).
- Hanssen, R. F., Weckwerth, T. M., Zebker, H. A. & Klees, R. High-resolution water vapor mapping from interferometric radar measurements. *Science* **283**, 1297–1299 (1999).
- White, D. E., Hutchinson, R. A. & Keith, T. E. C. The geology and remarkable thermal activity of Norris Geyser Basin, Yellowstone National Park, Wyoming. *US Geol. Surv. Prof. Pap.* **1456**, 1–84 (1988).
- Ingebritsen, S. E. & Rojstaczer, S. A. Controls on geyser periodicity. *Science* **262**, 889–892 (1993).
- Husen, S., Taylor, R., Smith, R. B. & Heasler, H. Changes in geyser eruption behaviour and remotely triggered seismicity in Yellowstone National Park produced by the 2002 $M 7.9$ Denali fault earthquake, Alaska. *Geology* **32**, doi:10.1130/G20381.1 (2004).
- Massonnet, D. & Feigl, K. L. Radar interferometry and its application to changes in the Earth's surface. *Rev. Geophys.* **36**, 441–500 (1998).
- Toda, S., Stein, R. S., Reasenberg, P. A. & Dieterich, J. H. Stress transferred by the $M_w = 6.9$ Kobe, Japan, shock: effect on aftershocks and future earthquake probabilities. *J. Geophys. Res.* **103**, 24543–24565 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Christiansen, S. Hurlwitz, J. Lowenstern, S. Ingebritsen, J. Savage and G. Waite for discussions, reviews and comments. European Space Agency (ESA) ERS-2 data were acquired through an ESA ENVISAT AO and through the WInSAR consortium (supported by NASA, NSF and the USGS). Help from K. Wendt and F. Boler with the GPS data is greatly appreciated. Collection of the GPS data, archived at UNAVCO, was funded by an NSF grant to R. Smith (Univ. of Utah).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.W.W. (cwicks@usgs.gov).

LETTERS

Early maize agriculture and interzonal interaction in southern Peru

Linda Perry¹, Daniel H. Sandweiss^{2,3}, Dolores R. Piperno^{1,4}, Kurt Rademaker³, Michael A. Malpass⁵, Adán Umire⁶ & Pablo de la Vera⁷

Over the past decade, increasing attention to the recovery and identification of plant microfossil remains from archaeological sites located in lowland South America has significantly increased knowledge of pre-Columbian plant domestication and crop plant dispersals in tropical forests and other regions^{1–4}. Along the Andean mountain chain, however, the chronology and trajectory of plant domestication are still poorly understood for both important indigenous staple crops such as the potato (*Solanum* sp.) and others exogenous to the region, for example, maize (*Zea mays*)^{5,6}. Here we report the analyses of plant microremains from a late preceramic house (3,431 ± 45 to 3,745 ± 65 ¹⁴C BP or ~3,600 to 4,000 calibrated years BP) in the highland southern Peruvian site of Waynuna. Our results extend the record of maize by at least a millennium in the southern Andes, show on-site processing of maize into flour, provide direct evidence for the deliberate movement of plant foods by humans from the tropical forest to the highlands, and confirm the potential of plant microfossil analysis in understanding ancient plant use and migration in this region.

Test excavations at the open-air site of Waynuna (15° 16' 01" S, 72° 44' 55" W; site CO-38; ref. 7) in the Cotahuasi Valley, Arequipa Department, southern Peru, uncovered a portion of a late preceramic house (Fig. 1 and Table 1). The site lies in the shadow of Cerro Aycano, a peak at 4,475 m above sea level (a.s.l.) that anchors the northwest end of the Alca obsidian source, one of the two most important sources of obsidian in the south Central Andes⁸ and a magnet for humans since Paleoindian times⁹. Waynuna occupies a natural terrace on a narrow finger ridge, which affords the site both morning and afternoon sun; other natural terraces are within an hour's walk and the site is less than 500 m from a perennial stream that descends from Cerro Aycano. At 3,625 m a.s.l., Waynuna lies at the interface of two distinct ecological zones with different agricultural potentials that currently support dissimilar staple crops (for example, see refs 10, 11). From 2,300 to 3,600 m a.s.l. is the *qheshwa* or *valle* temperate zone, characterized by intensively irrigated maize production. From 3,600 to 4,000 m a.s.l. is the *suní* or *echadero*, a cool upland zone devoted to tuber cultivation that is beyond the practical limits for irrigation-dependent maize farming.

Large surface areas of Waynuna are today occupied by abandoned agricultural terraces apparently built during the Middle Horizon (~1,200–950 calibrated years (cal yr) BP; ref. 7), indicating that the site itself was once under cultivation, albeit at the modern upper limit for maize. The preceramic deposits at Waynuna were covered by 1 m of Middle Horizon terrace fill containing abundant faunal remains, lithics (largely obsidian), potsherds, and several neonatal or infant

burials. The preceramic strata yielded a lower density of artefacts, although obsidian debitage remained abundant. Level 3a–3b was a thick homogenous level, possibly fill, that included three small fragments of grinding stones. Level 4, the use deposit, covered a well-defined dirt floor (level 5) inside an arched house wall formed by boulders up to 50-cm high (Fig. 1). Although only about a quarter of the house was excavated, the plan appears circular with a diameter of about 2.10 m. A 30-cm opening on the north side served as a doorway. This layout is typical of Late Preceramic period houses in the Central Andes, although the diameter falls on the low end of the known range¹². Below the floor were two stratigraphically equivalent, ashy levels (6 and 7i), separated by a protrusion of the sterile surface (level 7). Some archaeological material had intruded into this sterile soil, but density dropped off rapidly with depth while the matrix remained unchanged.

Waynuna is in a region of seasonal rainfall, and dead wood decomposes rapidly on the soil surface. Site elevation is above the modern limit for large trees; thus, available wood does not contain core material significantly older than the date of tree death. For both of these reasons, our bulk charcoal samples should not overestimate site age. Furthermore, the dates from Waynuna are coeval with the end of the Late Preceramic period in the south Central Andes; thus, it is unlikely that the true age of the site is younger than the age indicated by the ¹⁴C results. Radiocarbon dates for levels 4–7i (those on and below the floor) are clustered, and level 3b seems to be slightly more recent (Table 1).

Because plant macroremains preserve poorly in this seasonally wet environment, microbotanical analysis of starches and silica phytoliths was selected as the best approach for examining the archaeological assemblage. Unscreened soil samples from all preceramic strata from level 3b down to and including the preoccupation surface (level 7), and three unwashed fragments of grinding stones from levels 3a and 3b were analysed for plant microfossil remains by L.P. and D.R.P. by standard methods (see Supplementary Information). We recovered 1,077 starch granules from the soil samples and lithic tools (Table 2). Maize remains (Fig. 2d, e) were the most prevalent: we found 970 definitively identified starch granules and 49 probable maize granules. Remains of arrowroot (*Maranta* sp.; Fig. 2b) were recovered from level 3b as starch from a groundstone tool fragment. Arrowroot is a plant of the lowland tropical forest and cannot be grown above about 1,000 m a.s.l. (ref. 1). The starchy remains of what is probably potato (cf. *Solanum* sp.) occurred on a single groundstone fragment also from level 3b.

Phytolith content of the samples was generally good, apart from a sample collected from a culturally sterile context (level 7) where

¹Archaeobiology Program, Department of Anthropology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, MRC 112, Washington, DC 20013-7012, USA. ²Department of Anthropology, S. Stevens Hall, University of Maine, Orono, Maine 04469-5773, USA. ³Climate Change Institute, Bryand Global Sciences Center, University of Maine, Orono, Maine 04469, USA. ⁴Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Balboa, Ancon, Republic of Panama. ⁵Department of Anthropology, G121 Gannett Center, Ithaca College, 953 Danby Road, Ithaca, New York 14850, USA. ⁶Museo Contisuyo, Jr. Tacna 294, Moquegua, Perú. ⁷Instituto Nacional de Cultura, Alameda San Lázaro 120, Arequipa, Perú.

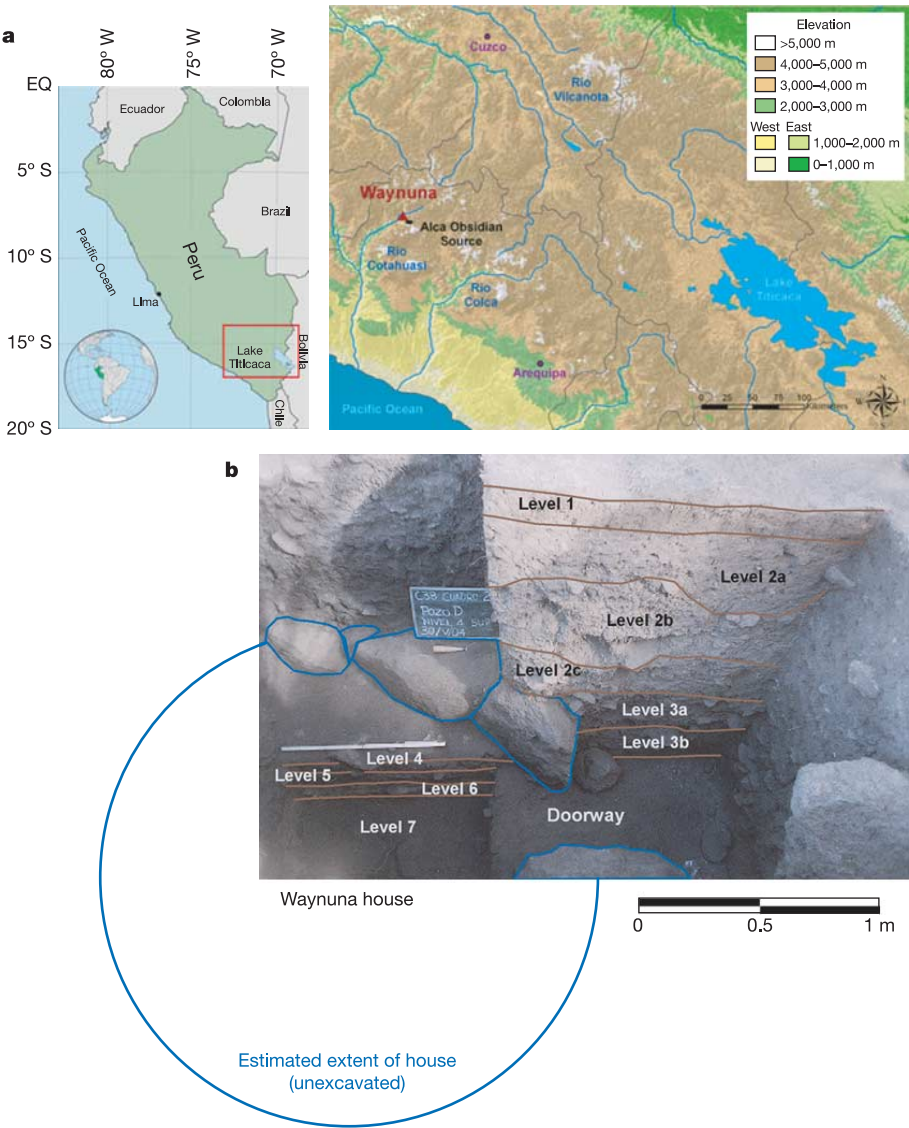


Figure 1 | Location of Waynuna and detail of the excavated Late Preceramic house and associated stratigraphy. **a**, Location of Waynuna. **b**, West profile of Waynuna house showing strata boundaries, pit 1A (right), pit 1C (lower left), and unmarked upper profile of pit 2D in the left background. Outlined stones are part of the preceramic house, the reconstructed outline of which is shown in blue.

virtually no phytoliths occurred (Table 2). Both maize leaf and cob phytoliths were recovered from level 5 (catalogue no. 33, second sample; Fig. 2c). Cob phytoliths were present in most other samples but, with the exception of those from level 5, maize leaf silica bodies could not be unequivocally identified because cross-bodies were very

few in number ($n < 10$ from scans of the whole slide) or were of a size class smaller than those that identify maize. In level 3b (catalogue no. 30), conical phytoliths diagnostic of the plant family Marantaceae were found (Fig. 2a). They are typical of silica bodies that occur in the bracts and flesh of arrowroot rhizomes, and can also be found in

Table 1 | Radiocarbon dates from Waynuna

Cat. level	Sq	Pit	Feat.	RCY BP	Lab. No.	cal yr BP 2σ	δ ¹³ C	Comments
15: 3b	1	C		3,431 ± 45	BGS 2576	3,472–3,723, 3,797–3,816	–24.00‰ (+16 yr)	Posthouse deposit
16: 3b	1	C	1C-1	3,544 ± 55	BGS 2578	3,590–3,600, 3,612–3,905	–24.18‰ (+13 yr)	Posthouse deposit
18: 4	1	C		3,628 ± 65	BGS 2552	3,690–4,010, 4,028–4,083	–24.83‰ (+3 yr)	Use deposit on floor
32: 4	2	D		3,545 ± 45	BGS 2577	3,638–3,886	–23.97‰ (+16 yr)	Use deposit on floor
20: 4	1	C	1C-2	3,582 ± 100	BGS 2581	3,559–4,090, 4,131–4,137	–24.47‰ (+9 yr)	Feature on floor (L. 5)
22: 5	1	C		3,631 ± 45	BGS 2575	3,716–3,986, 4,049–4,063	–24.40‰ (+10 yr)	Floor
33: 5	2	D		3,509 ± 70	BGS 2580	3,486–3,502, 3,506–3,524, 3,555–3,901	–25.08‰ (–1 yr)	Floor
34: 6	2	D		3,745 ± 65	BGS 2573	3,842–4,183, 4,196–4,234	–24.80‰ (+3 yr)	Subfloor deposit
24: 7i	1	C		3,590 ± 45	BGS 2551	3,689–3,931, 3,942–3,969	–24.79‰ (+3 yr)	Subfloor deposit
25: 7	1	C		3,633 ± 85	BGS 2574	3,640–3,672, 3,677–4,093, 4,125–4,143	–23.69‰ (+21 yr)	Sterile with intruded material from 6, 7i

All assays were run at the Jaan Terasmae Radiocarbon Laboratory, Brock University, Canada, on bulk charcoal samples corrected for ¹²C/¹³C isotopic fractionation (δ¹³C = [(¹³C/¹²C)_{sample}/(¹³C/¹²C)_{standard}] – 1). Calibrations were run with Calib 5.0.1 (refs 28–30) using the Southern Hemisphere atmosphere correction. Cat., catalogue no.; Feat., feature; RCY, radiocarbon years (±s.d.).

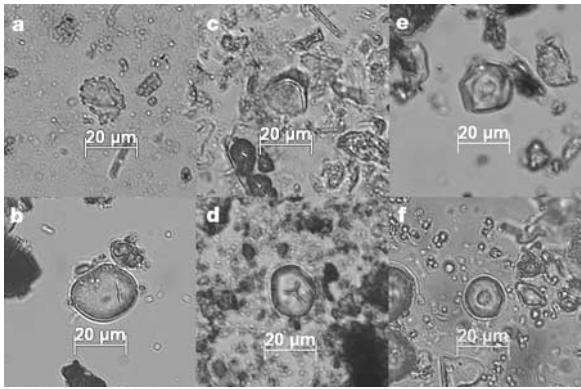


Figure 2 | Starch granules and phytoliths recovered from sediment samples and tool surfaces. **a**, Phytolith of Marantaceae (catalogue no. 30). **b**, Starch granule of arrowroot (tool 11). **c**, Rondel phytolith of maize (catalogue no. 33 segunda). **d**, Starch granule of maize (catalogue no. 32). **e**, Damaged starch granule of maize showing a scooped out area at the centre (tool 29). **f**, Modern starch granule subjected to grinding.

other species of the Marantaceae¹³. Given the occurrence of starch remains from only a single member of the Marantaceae, arrowroot, it is likely the phytoliths derive from this cultivar. They provide additional evidence for the presence of tropical forest cultivars in this highland location. The possibility that the microfossils are from wild members of the Marantaceae can be ruled out because this lowland family is not naturally distributed at this elevation.

Overall, the total numbers of microfossils remain fairly consistent throughout the column, with the exception of a peak in concentrations of both starches and phytoliths at level 5, the stratigraphic level that corresponds with the house floor. This finding supports the functional identification of the floor.

We divided maize granules into categories on the basis of both morphological features and structural damage resulting from grinding and pounding (Table 2 and Supplementary Information). Starch granules from flour-type corns differ from harder endosperm corns such as popcorn or dent corn, and this distinction has been noted in archaeological starch residues from Ecuador, Panama and Venezuela^{3,14–16}. Although many modern Peruvian maize specimens in our comparative collections contain both morphologies within the same kernel, the presence of two distinct assemblages on the lithic tools, each dominated by either flour or hard endosperm type starch, indicate that at least two races of maize were in use at Waynuna (Table 2).

The percentage of maize starch granules that showed damage consistent with that produced by grinding ranged from 11 to 51% in each sample (Fig. 2e, f). The nature of maize starch itself supports the hypothesis that the observed damage is due to grinding. The surface of maize starch granules is pockmarked by small pores¹⁷. Enzymes enter the granule through these pores so that the granule is effectively digested from the inside out^{17,18}. Thus, surface damage would not result from the digestive enzymes of soil-borne microorganisms but from external mechanical forces such as grinding.

Although many models of the rise of the state in ancient South America have relied on intensive agriculture, often involving maize (for example, see refs 5, 19, 20), the documented history of Andean maize is discontinuous. To the best of our knowledge, the earliest conclusive evidence for maize previously reported from the Central Andean highlands dates to ~2,500 cal yr BP (ref. 5). This region saw the rise of the maize-dependent empires of the Wari, Tiwanaku and Inca²¹. The starch and phytolith assemblages from Waynuna testify to the use of maize by 4,000 cal yr BP, whereas damage to starch granules extracted from the worn surfaces of the grindstone tool fragments confirms on-site processing of maize for food use. The presence of both leaf and cob phytoliths in the same context strongly suggests that maize was cultivated and processed on site.

Potato is one of the most important members of the Andean root crop complex and it can be cultivated at this elevation^{10,22}. Potato starch granules have been isolated and studied from tubers preserved on and near the arid Peruvian coast at sites dating to between ~4,000 cal yr BP and 3,000 cal yr BP (ref. 22). Although at present we cannot positively determine whether the Waynuna residues are derived from a wild or domesticated species of *Solanum*, retrieval of the starches from this context indicates that they will survive when macroremains of the tubers have decayed.

Because the lowland domesticate arrowroot does not occur in the rich macrofossil assemblages of domesticated plants preserved in arid coastal Peruvian sites (for example, see refs 1, 22), this crop most probably arrived at Waynuna from the Amazonian rainforest to the east. The nearest interzonal access route passes by contemporaneous sites that received the lithic raw material obsidian from the Alca source near Waynuna⁸. The arrowroot findings at Waynuna are the earliest empirical evidence for movement of goods and/or people from eastern lowland zones into the Andean highlands. Such transfer of products between these ecological zones has been a cornerstone of models of highland Andean civilization from the Initial period and Chavin to the Inca empire (~3,800 to 418 cal yr BP), but until now we lacked direct, early evidence from the highlands on the eve of emerging complexity^{23,24}. At present, the evidence for other lowland crops (achira, peanuts, manioc) postulated to have been important

Table 2 | Starch and phytolith frequencies from sediment samples and tool fragments

Cat. level	Maranta sp.	Marantaceae phytoliths	cf. Solanum sp.	Zea mays cob phytoliths	Zea mays leaf phytoliths	Zea mays	Zea mays (soft)	Zea mays (hard)	Zea mays (damaged)	cf. Zea mays	Unidentified	Grand total
Soil samples												
30: 3b		P		-	A	5	14	42	22	11	3	97
31: El. 2D-2		A		P	A		6	15	25	3	2	51
32: 4		A		-	A		2	29	16		2	49
33: 5 primera		A		-	A	6	13	24	16	2	3	64
33: 5 segunda		A		C	P	1	50+	50+	50+	6	8	165+
34: 6		A		C	-	1	50+	9	16			76+
36: 7i		A		P	-	3	19	24	14	15	2	77
35: El. 2D-3		A		A	A	1	9	7	9	2	2	30
37: 7 (sterile)		A		A	A	1	6	16	16	6		45
Total (n = 9)	0	P		P	P	18	169	216	184	45	22	654
Lithic tools												
29: 3a							11	53	29	3	6	102
11: 3b	4						104	28	17	1	11	165
30: 3b			3				115	8	18		12	156
Total (n = 3)	4	-	3	-	-	0	230	89	64	4	29	423
Grand total	4	P	3	P	P	18	399	305	248	49	51	1,077

Numerals indicate the numbers of starch granules recovered. Letters indicate whether phytoliths were present (P), absent (A) or common (C) in samples.

items of exchange and/or symbolism among tropical forest and highland Andean peoples still only comes from iconographic depictions²⁴. Because these plants also produce diagnostic starch grains, future microfossil work should reveal more information.

Waynuna's topographic setting, proximity to perennial water, and position at the ecotone where both maize and root crop agriculture are productive, fit traditional criteria of Andean peasant farmers for a desirable village site²⁵. The archaeobotanical data indicate that this preference for settlement at the maize–tuber ecotone has considerable time depth in the region. The successful identification of starch granules damaged by grinding supports previous research²⁶, provides independent evidence for assessing stone tool function and plant processing strategy, and confirms that starch remains derived from sediments represent food residues. The successful recovery of the microremains of both seed and tuber crops from Waynuna, including starch granules and phytoliths from two different structures of the same food plant species, indicates that multiproxy microfossil analysis may be the most effective means for documenting crop plants and understanding their history and role in many ancient Andean contexts.

METHODS

Methods for starch extractions from soils²⁷ were modified from those published elsewhere, and standard procedures were used for analyses of lithic tools^{14,15}. Extraction methods for phytolith analysis of soil samples followed standard procedures¹³. Microfossil identifications were based on modern reference collections of more than 2,000 Neotropical and Andean plant species housed at the Smithsonian Tropical Research Institute in Panama and the Department of Anthropology, National Museum of Natural History in Washington DC. For detailed methods, see Supplementary Information.

Received 23 May; accepted 29 September 2005.

- Piperno, D. R. & Pearsall, D. M. *The Origins of Agriculture in the Lowland Neotropics* (Academic, San Diego, 1998).
- Piperno, D. R. & Stothert, K. E. Phytolith evidence for early holocene Cucurbita domestication in southwest Ecuador. *Science* **299**, 1054–1057 (2003).
- Pearsall, D. M., Chandler-Ezell, K. & Zeidler, J. A. Maize in ancient Ecuador: results of residue analysis of stone tools from the Real Alto site. *J. Arch. Sci.* **31**, 423–442 (2004).
- Iriarte, J. et al. Evidence for cultivar adoption and emerging complexity during the mid-Holocene in the La Plata basin. *Nature* **432**, 614–617 (2004).
- Burger, R. L. & Van der Merwe, N. J. Maize and the origin of highland Chavin civilization: an isotopic perspective. *Am. Anthropol.* **92**, 85–95 (1990).
- Hastorf, C. A. in *The Prehistory of Food* (eds Gosden, C. & Hather, J. G.) 35–58 (Routledge, London, 1998).
- Jennings, J. *Prehistoric Imperialism And Cultural Development In The Cotahuasi Valley, Peru*. Dissertation, Univ. California Santa Barbara (2002).
- Burger, R. L., Mohr Chávez, K. L. & Chávez, S. J. Through the glass darkly: prehispanic obsidian procurement and exchange in southern Peru and northern Bolivia. *J. World Prehistory* **14**, 267–362 (2000).
- Sandweiss, D. H. et al. Quebrada Jaguay: early maritime adaptations in South America. *Science* **281**, 1830–1832 (1998).
- Pulgar Vidal, J. *Geografía del Perú* 9th edn (PEISA, Lima, 1987).
- Trawick, P. B. *The Struggle for Water in Peru: Comedy and Tragedy in the Andean Commons* (Stanford Univ. Press, Stanford, 2002).
- Malpass, M. A. & Stothert, K. E. Evidence for preceramic houses and household organization in western South America. *Andean Past* **3**, 137–163 (1992).
- Piperno, D. R. *Phytoliths: a Comprehensive Guide for Archaeologists and Paleoecologists* (AltaMira, Lanham, MD, In the press).
- Perry, L. Starch analyses reveal the relationship between tool type and function: an example from the Orinoco valley of Venezuela. *J. Arch. Science* **31**, 1069–1081 (2004).
- Perry, L. Reassessing the traditional interpretation of 'manioc' artifacts in the Orinoco Valley of Venezuela. *Latin Am. Ant.* **16**, 409–426 (2005).
- Piperno, D. R., Ranere, A. J., Holst, I. & Hansell, P. Starch grains reveal early root crop horticulture in the Panamanian tropical forest. *Nature* **407**, 894–897 (2000).
- Fannon, J. E., Hauber, R. J. & BeMiller, J. N. Surface pores of starch granules. *Cereal Chem.* **69**, 284–288 (1992).
- Leach, H. W. & Schoch, T. J. Structure of the starch granule II. Action of various amylases on granular starches. *Cereal Chem.* **38**, 34–46 (1961).
- Haas, J., Pozorski, S. & Pozorski, T. (eds) *The Origins and Development of the Andean State* (Cambridge Univ. Press, Cambridge, 1987).
- Wilson, D. J. *Prehispanic Settlement Patterns in the Lower Santa Valley, Peru* (Smithsonian Institution Press, Washington DC, 1988).
- Moseley, M. E. *The Incas and Their Ancestors* 2nd edn (Thames & Hudson, London and New York, 2001).
- Ugent, D., Pozorski, S. & Pozorski, T. Archaeological potato tuber remains from the Casma Valley of Peru. *Econ. Bot.* **36**, 182–192 (1982).
- Lathrap, D. W. in *Variation in Anthropology: Essays in Honor of John C. McGregor* (eds Lathrap, D. W. & Douglas, J.) 91–105 (Illinois Archaeological Survey, Urbana, 1973).
- Lathrap, D. W. The antiquity and importance of long distance trade relationships in the moist tropics of pre-Columbian South America. *World Arch.* **5**, 170–186 (1973).
- Brush, S. *Mountain Field and Family: The Economy and Human Ecology of an Andean Valley* (Univ. Pennsylvania Press, Philadelphia, 1977).
- Babot, M. del P. in *Phytolith and Starch Research in the Australian-Pacific-Asian Regions: the State of the Art* (eds Hart, D. M. & Wallis, L. A.) 69–81 (Pandanus Books, Canberra, 2003).
- Fullagar, R., Loy, T. & Cox, S. in *A Closer Look: Recent Australian Studies of Stone Tools* (ed. Fullagar, R.) 49–60 (Archaeological Computing Laboratory, Univ. Sydney, Sydney, 1998).
- McCormac, F. G. et al. SHCal04 Southern Hemisphere calibration 0–11.0 cal kyr BP. *Radiocarbon* **46**, 1087–1092 (2004).
- Reimer, P. J. et al. IntCal04 terrestrial radiocarbon age calibration, 0–26 cal kyr BP. *Radiocarbon* **46**, 1029–1058 (2004).
- Stuiver, M. & Reimer, P. J. Extended ¹⁴C data base and revised CALIB 3.0 ¹⁴C age calibration program. *Radiocarbon* **35**, 215–230 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements D. Goldstein provided modern Andean plant materials for comparative purposes. This work was supported by a grant from the Heinz Charitable Trust Latin American Archaeology Program, and funding from FERCO and the Office of the Provost, Ithaca College, and from the Smithsonian Tropical Research Institute and the Smithsonian National Museum of Natural History.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.P. (perryli@si.edu).

LETTERS

Coral reef diversity refutes the neutral theory of biodiversity

Maria Dornelas^{1,2}, Sean R. Connolly^{1,2} & Terence P. Hughes¹

The global decline of coral reefs^{1,2} highlights the need to understand the mechanisms that regulate community structure and sustain biodiversity in these systems. The neutral theory, which assumes that individuals are demographically identical regardless of species, seeks to explain ubiquitous features of community structure and biodiversity patterns^{3–5}. Here we present a test of neutral-theory predictions with the use of an extensive species-level data set of Indo-Pacific coral communities. We show that coral assemblages differ markedly from neutral-model predictions for patterns of community similarity and the relative abundance of species. Within local communities, neutral models do not fit relative abundance distributions as well as the classical log-normal distribution. Relative abundances of species across local communities also differ markedly from neutral-theory predictions: coral communities exhibit community similarity values that are far more variable, and lower on average, than the neutral theory can produce. Empirical community similarities deviate from the neutral model in a direction opposite to that predicted in previous critiques of the neutral theory^{6–9}. Instead, our results support spatio-temporal environmental stochasticity as a major driver of diversity patterns on coral reefs^{10,11}.

The neutral theory was initially proposed to explain the dynamics of communities that sustain high biodiversity whose species compete for space by a lottery mechanism. This set of assumptions seems, at least superficially, to be better suited to coral communities than to most other animal assemblages¹². Testing the neutral theory against coral biodiversity patterns is therefore essential in assessing its generality and can provide much insight into mechanisms of biodiversity maintenance in these increasingly threatened ecosystems. Most tests of the neutral theory have focused on assessing the fit of a neutral model to relative abundance patterns within local communities, or for a metacommunity as a whole^{4,5,13}. However, it has recently become clear that such approaches, used in isolation, provide only weak tests of community structure models^{14,15}. Therefore, assessing the predictions of a theory at multiple levels¹⁶ is preferable, particularly when competing theories make different predictions. In neutral communities, variability in relative abundances between species is due entirely to demographic stochasticity or 'ecological drift'. Theoretical analyses of neutral models have shown that this process leads to divergence between communities¹⁷, and thus community similarity is determined by the strength of dispersal limitation^{4,17,18}. In contrast, niche apportionment mechanisms have been invoked to argue that higher levels of community similarity should be observed under niche assembly than under neutral dynamics^{6–9}. These contrasting predictions provide an ideal opportunity to test the neutral model against empirical data.

Here we use an extensive data set of species-level abundances on coral reefs to conduct a multiscale test of the hypothesis that the structure of coral assemblages is consistent with the neutral theory. Using neutral-model parameters fitted to abundance distributions

within local communities, we predicted pairwise community similarity statistics for a network of local communities. The predicted similarities are consistently high (Fig. 1a). In contrast, the observed similarities of coral communities are markedly lower, and more variable, than predicted by the neutral theory (Fig. 1b). Furthermore, the neutral model's fit to abundance distributions within local communities is significantly worse than the fit of a Poisson log-normal distribution (model selection supports the Poisson log-normal with more than 99% confidence; Supplementary Fig. S1 and Supplementary Methods). Hence, neither community similarity patterns nor local abundance distributions of reef corals are consistent with neutral-model predictions.

Our finding that observed community similarities are lower than neutral-model predictions is unexpected, and is contrary to the widespread view that high similarity over space and time in empirical data constitutes evidence against the neutral theory^{6–9}. However,

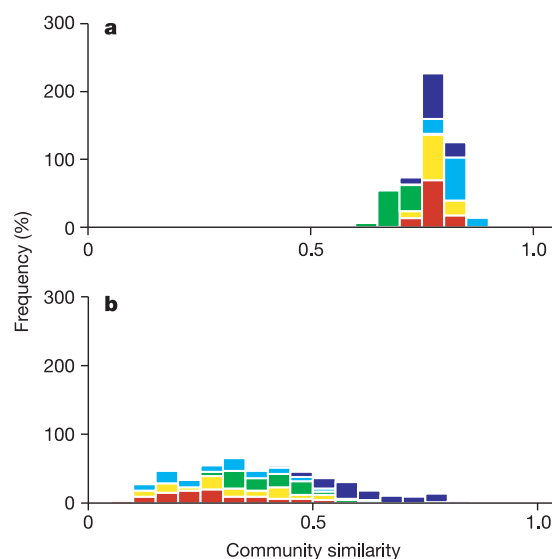


Figure 1 | Frequency distributions of Bray-Curtis similarities.

a, Distributions for local communities simulated with parameters estimated from the data's species abundance distributions; **b**, distributions for observed coral assemblages on reef slopes. Parameter estimates were $m = 0.905, 0.916, 0.915, 0.867$ and 0.748 and $\theta = 25.4, 26.5, 30.6, 10.8$ and 6.3 for Indonesia (red), Papua New Guinea (yellow), Solomon Islands (green), Samoa (light blue) and French Polynesia (dark blue), respectively. The heights of the bars sum to 100% separately for each region, so the stacked bars may exceed 100%. Note the low mean and high variance of the observed distributions in comparison with those of the simulations. Flat and crest assemblages differ from neutral-model predictions in a very similar fashion (Fig. 2).

¹ARC Centre of Excellence for Coral Reef Studies and ²School of Marine Biology and Aquaculture, James Cook University, Townsville, Queensland 4811, Australia.

others have argued that the neutral theory can produce high similarities^{4,18}. Therefore, to achieve a better understanding of community similarity under the neutral theory, we conducted a comprehensive investigation of neutral-model dynamics. Our results confirm that the neutral theory can produce extremely high similarities, provided that immigration rates from the metacommunity are high, or diversity is very low (Fig. 2, Supplementary Fig. S2 and Supplementary Methods). However, our analyses also reveal that variance in community similarity under neutral dynamics is consistently low, much lower than in our data, regardless of parameter values (Fig. 2, Supplementary Fig. S3 and Supplementary Methods), thereby also confirming that this large discrepancy between neutral-model predictions and our data are robust to uncertainty in parameter estimates.

The higher variance of real coral communities than in neutral-theory predictions indicates a strong role for environmental variability in determining patterns of community similarity on coral reefs. This high variance in community similarity indicates that the data contain a much broader mixture of sites that are very similar and very different, in comparison with the neutral theory. Because habitat differences were minimized by our study design (see Methods), the most likely source of this variability is spatio-temporal environmental stochasticity: the tendency for different local communities to experience fluctuations in environmental conditions differently. On coral reefs, even adjacent reefs can have markedly different environmental histories¹⁹. Such historical differences would tend to decrease mean similarity if species differ in their responses to environmental fluctuations. These mechanisms would also be expected to increase the variability in community similarity values: communities that have experienced similar environmental histories would tend to have a similarity that is higher than average (for example, communities at similar successional stages), whereas communities with markedly different histories would tend to have a similarity that is lower than average. In contrast, neutral models assume that metacommunities are environmentally homogeneous in space and time. Indeed, differences between species in responses to environmental fluctuations challenge a core assumption of the neutral theory: that species are demographically identical.

In addition to the assumptions of environmental homogeneity

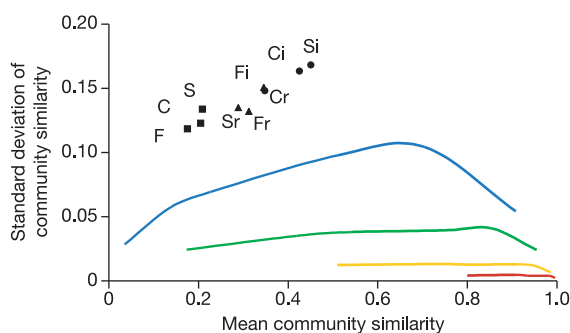


Figure 2 | Bray-Curtis similarity distributions for neutral simulations and coral communities. Because community similarity frequency distributions are approximately normal (Supplementary Information), they can be compared in terms of their mean and standard deviation. Each contour plots the mean and standard deviation of community similarity for $1 < \theta < 500$ (from right to left on each contour) for a particular value of m (0.999 in red, 0.1 in yellow, 0.01 in green and 0.001 in blue). Bray-Curtis similarity distributions for coral assemblages in each habitat are plotted as black points: Si within island slope assemblages, Ci within island crest assemblages, Fi within island flat assemblages, Sr within region slope assemblages, Cr within region crest assemblages, Fr within region flat assemblages, S among all slope assemblages, C among all crest assemblages, and F among all flat assemblages. Note how all data points, regardless of spatial scale or habitat, fall outside and comparably distant from the contours generated by neutral simulations.

and neutrality, which are shared by all neutral models²⁰, there are other assumptions that differ between alternative formulations. We believe these assumptions are less plausible explanations for our results. First, neutral models can have different assumptions about how speciation occurs⁴. However, because speciation occurs rarely, in comparison with changes in relative abundance within species, it seems unlikely that different speciation mechanisms would have a marked effect on community similarity patterns, at least for realistic speciation rates²¹. A second assumption is the 'mainland-island' framework: metacommunities are held constant during local community dynamics⁴. An alternative formulation uses an archipelago framework, with the metacommunity changing as local communities change³. Although this might decrease community similarity over time, there is no obvious reason to expect it to increase variance in similarity between local communities, which is a major cause of the discrepancies between coral communities and neutral-model predictions. Last, spatially explicit neutral models predict decreasing similarity with distance^{4,22}. Given that our most distant sites are more than 10,000 km apart, dispersal limitation is certainly occurring within our sampling scale. If dispersal limitation were the principal cause of the discrepancies between our data and the neutral model, the empirical data should converge towards the neutral model (that is, they should move down and to the right on Fig. 2) as the spatial scale at which community similarity is calculated decreases. However, the data do not converge to the model: the high variance in coral community similarity does not decrease as the spatial scale decreases, even when similarity is calculated only for sites less than 10 km apart (Fig. 2).

Because coral communities have been viewed as being among the assemblages most likely to exhibit neutral dynamics^{12,23–25}, our findings challenge the neutral theory's utility as a general theory of biodiversity and biogeography. The neutral theory was initially proposed as a null model for macroecology, predicting the diversity patterns that could arise from the action of demographic stochasticity and dispersal limitation alone⁴. However, early findings that neutral models can exhibit a close fit to empirical species–abundance distributions rapidly shifted the focus towards whether or not the neutral theory is sufficient, by itself, to explain macroecological regularities, such as species–abundance distributions and species–area relationships^{5,13,16,26}. Although our results do not support the latter hypothesis, they do show how using the neutral theory as a null model can shed light on the mechanisms responsible for macroecological patterns.

In recent years there has been a renewed focus on the use of ecological theory to inform conservation priorities. At the large ecological scales relevant to this endeavour, the prevailing view is that niche apportionment rules stabilize community structure over space and time^{6–9,27}. However, our results support an alternative view in which species' different responses to spatial and temporal environmental fluctuations are crucial in the maintenance of biodiversity¹⁰. From this perspective, protecting biodiversity requires preserving the patterns of connectivity that allow species to find and exploit suitable environments that are patchy over both space and time. Fluctuation-mediated coexistence has received comparatively little attention in the continuing controversy about the neutral theory, despite the existence of tractable models of the relative abundance of species that explicitly incorporate effects of environmental stochasticity¹¹. Given the accelerating pace of coral reef habitat loss², a renewed focus on these mechanisms and on the body of ecological theory associated with them is urgent.

METHODS

Sampling followed a hierarchical design with three distinct spatial scales: site, island and region²⁸. The five regions encompass a threefold gradient in coral species richness²⁹. There were three islands within each region, four sites within each island, and three local communities at each site: one from each of three reef habitats (flat, crest and slope). Within each local community, ten random

transects 10 m in length were established, and the colonies intercepted by each transect were counted and identified to species.

We estimated the parameters of the neutral model, θ and m (Supplementary Methods), by fitting the predicted species abundance distribution (the 'zero-sum multinomial' or ZSM) to each of the local coral communities, using both simulation-based⁴ and analytical methods^{5,20}. The fit of the ZSM was compared with the log-normal distribution using Akaike's information criterion¹⁴ (Supplementary Methods). We used parameters estimated from fits to local community abundance distributions to simulate the dynamics of 15 metacommunities: one for each habitat type in each region (Supplementary Methods). Bray–Curtis community similarities³⁰ were then calculated for all pairs of simulated local communities within the same metacommunity; these similarities were compared with the distribution of Bray–Curtis statistics for the empirical data. To comprehensively characterize the similarity patterns produced by the neutral theory, we also analysed distributions of community similarity for an exhaustive spectrum of values of θ and m (Supplementary Methods).

Received 30 August; accepted 19 December 2005.

- Bellwood, D. R., Hughes, T. P., Folke, C. & Nystrom, M. Confronting the coral reef crisis. *Nature* **429**, 827–833 (2004).
- Hughes, T. P. et al. Climate change, human impacts, and the resilience of coral reefs. *Science* **301**, 929–933 (2003).
- Bell, G. The distribution of abundance in neutral communities. *Am. Nat.* **155**, 606–617 (2000).
- Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, 2001).
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035–1037 (2003).
- Clark, J. S. & McLachlan, J. S. Stability of forest biodiversity. *Nature* **423**, 635–638 (2003).
- Gilbert, B. & Lechowicz, M. J. Neutrality, niches, and dispersal in a temperate forest understory. *Proc. Natl Acad. Sci. USA* **101**, 7651–7656 (2004).
- Pitman, N. C. A. et al. Dominance and distribution of tree species in upper Amazonian terra firme forests. *Ecology* **82**, 2101–2117 (2001).
- Terborgh, J., Foster, R. B. & Nunez, P. Tropical tree communities: A test of the nonequilibrium hypothesis. *Ecology* **77**, 561–567 (1996).
- Chesson, P. & Huntly, N. Short-term instabilities and long-term community dynamics. *Trends Ecol. Evol.* **4**, 293–298 (1989).
- Engen, S. & Lande, R. Population dynamic models generating the lognormal species abundance distribution. *Math. Biosci.* **132**, 169–183 (1996).
- Whitfield, J. Ecology: Neutrality versus the niche. *Nature* **417**, 480–481 (2002).
- McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature* **422**, 881–885 (2003).
- Connolly, S. R., Hughes, T. P., Bellwood, D. R. & Karlson, R. H. Community structure of corals and reef fishes at multiple scales. *Science* **309**, 1363–1365 (2005).
- McGill, B. J. Strong and weak tests of macroecological theory. *Oikos* **102**, 679–685 (2003).
- Adler, P. B. Neutral models fail to reproduce observed species-area and species-time relationships in Kansas grasslands. *Ecology* **85**, 1265–1272 (2004).
- Maurer, B. A. & McGill, B. J. Neutral and non-neutral macroecology. *Basic Appl. Ecol.* **5**, 413–422 (2004).
- Volkov, I., Banavar, J. R., Maritan, A. & Hubbell, S. P. The stability of forest biodiversity. *Nature* **427**, 696 (2004).
- Connell, J. H. et al. A long term study of competition and diversity of corals. *Ecol. Monogr.* **74**, 179–210 (2004).
- McGill, B., Maurer, B. A. & Weiser, M. D. Empirical evaluation of neutral theory. *Ecology* (in the press).
- Ricklefs, R. E. A comment on Hubbell's zero-sum ecological drift model. *Oikos* **100**, 185–192 (2003).
- Chave, J. & Leigh, E. G. A spatially explicit neutral model of beta-diversity in tropical forests. *Theor. Popul. Biol.* **62**, 153–168 (2002).
- Chave, J. Neutral theory and community ecology. *Ecol. Lett.* **7**, 241–253 (2004).
- Hubbell, S. P. A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs* **16**, s9–s21 (1997).
- Williamson, M. & Gaston, K. J. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *J. Anim. Ecol.* **74**, 409–422 (2005).
- Wootton, J. T. Field parameterization and experimental test of the neutral theory of biodiversity. *Nature* **433**, 309–312 (2005).
- Pandolfi, J. M. Limited membership in Pleistocene reef coral assemblages from the Huon Peninsula, Papua New Guinea: Constancy during global change. *Palaeobiology* **22**, 152–176 (1996).
- Karlson, R. H., Cornell, H. V. & Hughes, T. P. Coral communities are regionally enriched along an oceanic biodiversity gradient. *Nature* **429**, 867–870 (2004).
- Bellwood, D. R., Hughes, T. P., Connolly, S. R. & Tanner, J. Environmental and geometric constraints on Indo-Pacific coral reef biodiversity. *Ecol. Lett.* **8**, 643–651 (2005).
- Legendre, P. & Legendre, L. *Numerical Ecology* (Elsevier, Amsterdam, 1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Barbosa, H. Cornell, R. Karlson, staff, students and volunteers of the Centre for Coral Reef Biodiversity, and the High Performance Computing staff, at James Cook University for assistance with various stages of this project. This work was supported by James Cook University, the Australian Research Council, and the Fundação para a Ciência e a Tecnologia, Portugal.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.D. (maria.dornelas@jcu.edu.au).

Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK

Michael J. Tildesley¹, Nicholas J. Savill^{2,3}, Darren J. Shaw⁴, Rob Deardon^{2,5}, Stephen P. Brooks², Mark E. J. Woolhouse³, Bryan T. Grenfell^{6,7} & Matt J. Keeling¹

Foot-and-mouth disease (FMD) in the UK provides an ideal opportunity to explore optimal control measures for an infectious disease. The presence of fine-scale spatio-temporal data for the 2001 epidemic has allowed the development of epidemiological models that are more accurate than those generally created for other epidemics^{1–5} and provide the opportunity to explore a variety of alternative control measures. Vaccination was not used during the 2001 epidemic; however, the recent DEFRA (Department for Environment Food and Rural Affairs) contingency plan⁶ details how reactive vaccination would be considered in future. Here, using the data from the 2001 epidemic, we consider the optimal deployment of limited vaccination capacity in a complex heterogeneous environment. We use a model of FMD spread to investigate the optimal deployment of reactive ring vaccination of cattle constrained by logistical resources. The predicted optimal ring size is highly dependent upon logistical constraints but is more robust to epidemiological parameters. Other ways of targeting reactive vaccination can significantly reduce the epidemic size; in particular, ignoring the order in which infections are reported and vaccinating those farms closest to any previously reported case can substantially reduce the epidemic. This strategy has the advantage that it rapidly targets new foci of infection and that determining an optimal ring size is unnecessary.

The control of infectious diseases is often a compromise between the desire for large-scale implementation of control measures and what is logistically or economically feasible. Therefore, it is important that control measures are optimally targeted so as to minimize the adverse population-level impact of a disease. Optimal targeting will in general depend in a complex nonlinear fashion on disease dynamics, host demography and logistical constraints. Well-parameterized mathematical models of infectious disease spread are therefore a necessary tool for determining optimal control strategies, because a vast range of policies can be rapidly tested by simulation. One important aspect of optimal control is the balance between maintaining local control of outbreaks while attacking new foci of infection.

During the course of the 2001 FMD epidemic a range of control measures were implemented to try to reduce the transmission of infection; these included movement restrictions preventing long-range transportation of livestock, increased farm bio-security, rapid culling of livestock on infected premises (IPs) to limit virus excretion, and the reactive culling of livestock on 'at risk' farms to eliminate potential new cases. Reactively culled farms fell into two main

categories (see Methods): dangerous contacts (DCs) and contiguous premises (CPs). Reactive vaccination was considered during the later stages of the 2001 epidemic but was dismissed for several reasons (detailed in ref. 7), including concerns that the limited vaccination campaign that was logistically feasible at the time would have little impact on the epidemic.

The recently published European Union Directive⁸ suggests reactive vaccination as a preferred means of intervention should a foot-and-mouth disease outbreak occur in any member state. This is reflected in the UK by the new FMD contingency plan published by DEFRA⁶. However, neither document suggests a specific design for reactive vaccination programmes. Here, we concentrate on developing an optimal reactive and responsive vaccination programme based on the 2001 UK epidemic, assuming vaccination-to-live so that vaccinated animals do not have to be culled. Vaccination of cattle is combined with localized culling (following EU and DEFRA recommendations^{6,8}) in which all livestock on infected premises are culled within 24 hours of reporting and farms considered by epidemiological investigation to be at increased risk (DCs) are culled within 48 hours. By default, simulations do not include CP culling. The model closely follows that developed in refs 1 and 2; we use the farm census database from June 2000 to incorporate farm-level heterogeneities into our model while the epidemiological parameters are obtained by fitting to the 2001 spatio-temporal epidemic profile (see Methods and Supplementary Information).

We begin by assuming that vaccination (of cattle only) takes place within an annulus (defined by an inner and outer radius, R_{inner} and R_{outer}) around each IP^{3,9} such that all farms within a given distance of every IP will be identified for vaccination (see Methods). Vaccination of these farms is then prioritized in the order that the associated IPs were reported and, for those farms identified on the same day, priority is given to those furthest from the IP. Thus vaccination within each annulus is performed from the outside in. Protection of cattle is influenced by several factors. We assume a vaccine efficacy of 90% so that 10% of vaccinated cattle remain susceptible^{10,11}; in addition there is a four-day delay between vaccination and protection. Finally, although farms within an annulus are targeted for vaccination two days after the IP is reported, it may take longer before they are vaccinated owing to the logistical constraint that only a limited number of cattle can be vaccinated per day.

Varying R_{inner} and R_{outer} , we seek the vaccination annulus that minimizes the epidemic impact—defined as the number of farms that are infected (IPs) or culled as part of the control (DCs + CPs if appropriate). For a plausible vaccination capacity (defined as the

¹Department of Biological Sciences and Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK. ²Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. ³Epidemiology Group, Centre for Infectious Diseases, University of Edinburgh, Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh EH9 3JF, UK. ⁴Veterinary Clinical Studies, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, UK. ⁵Cambridge Infectious Diseases Consortium, Centre for Veterinary Science, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK. ⁶Center for Infectious Diseases Dynamics, Biology Department 208, Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA. ⁷Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA.

number of cattle that can be vaccinated per day), the model predicts that the optimal policy is to vaccinate in a complete ring with the inner radius set to zero (Fig. 1a). Extensive simulations have shown that this is true for all the situations examined here, indicating that—despite the delays involved—vaccination is still in general a locally effective control measure. This finding could be refined using predictive vaccination², where (on a case-by-case basis) the use of predictive simulations indicates that it may be optimal to ignore some local farms if they are likely to become infected before vaccination would confer immunity.

We focus on the behaviour of the model when the inner radius is fixed at zero (Fig. 1b and c), and observe several important patterns.

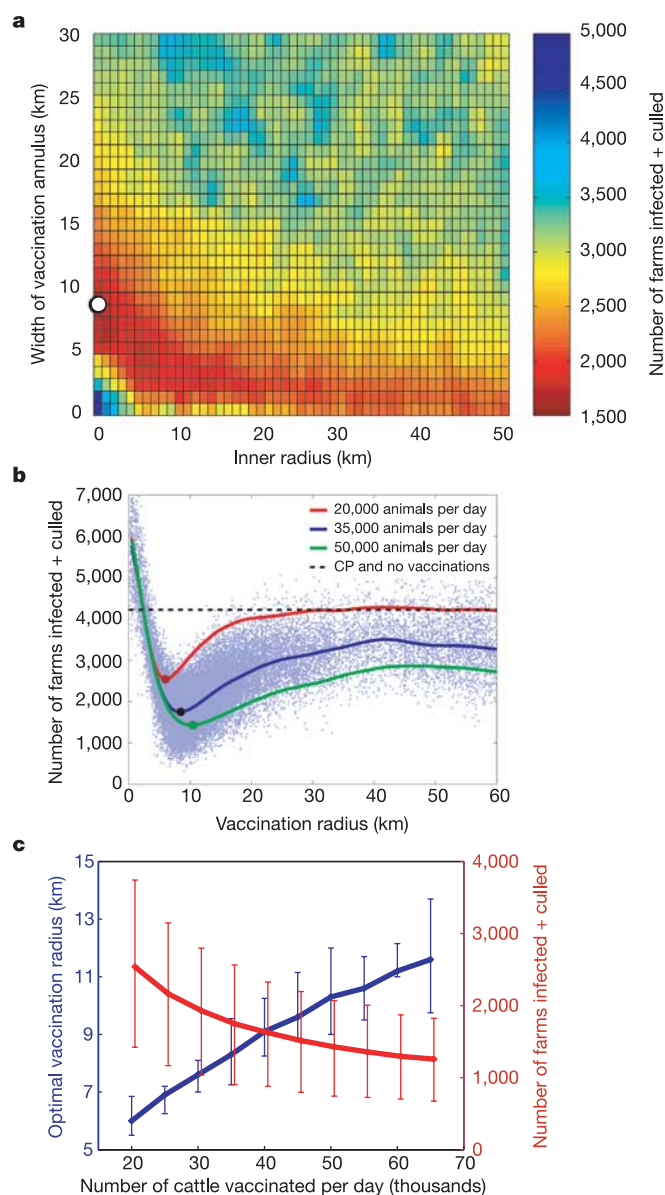


Figure 1 | Epidemic impact (number of farms infected and culled).

a, Epidemic impact as inner vaccination radius (R_{inner}) and width of the vaccination annulus ($R_{\text{outer}} - R_{\text{inner}}$) vary. The white dot marks the optimal strategy. **b**, Epidemic impact for various outer radii ($R_{\text{inner}} = 0$); points represent individual simulations, lines give a smoothed average for different vaccination capacities with large dots marking the optimal ring size. The dashed black line gives the average epidemic impact with prompt CP and DC culling but no vaccination. **c**, Epidemic impact (red) and optimal radius (blue) as the daily vaccination capacity is varied (with 95% confidence intervals for each point).

Increasing the vaccination capacity reduces the average epidemic impact (Fig. 1b and c), although there is considerable variation between simulations. The optimal vaccination ring increases with vaccination capacity. However, the number of animals within the optimal ring is not proportional to the number that can be vaccinated per day; instead it increases more rapidly, compensating for overlapping vaccination rings. This highlights the fact that calculation of the optimal ring size is a complex nonlinear problem that requires detailed spatial models. Finally, we note that optimal reactive ring vaccination at 35,000 animals per day (together with IP and DC culling) is a far better strategy, in terms of reducing the total number of farms lost, than a policy of CP culling (together with IP and DC culling but no vaccination) even if all the culls are performed promptly (Fig. 1b).

Armed with these insights, we then explore the effect of varying

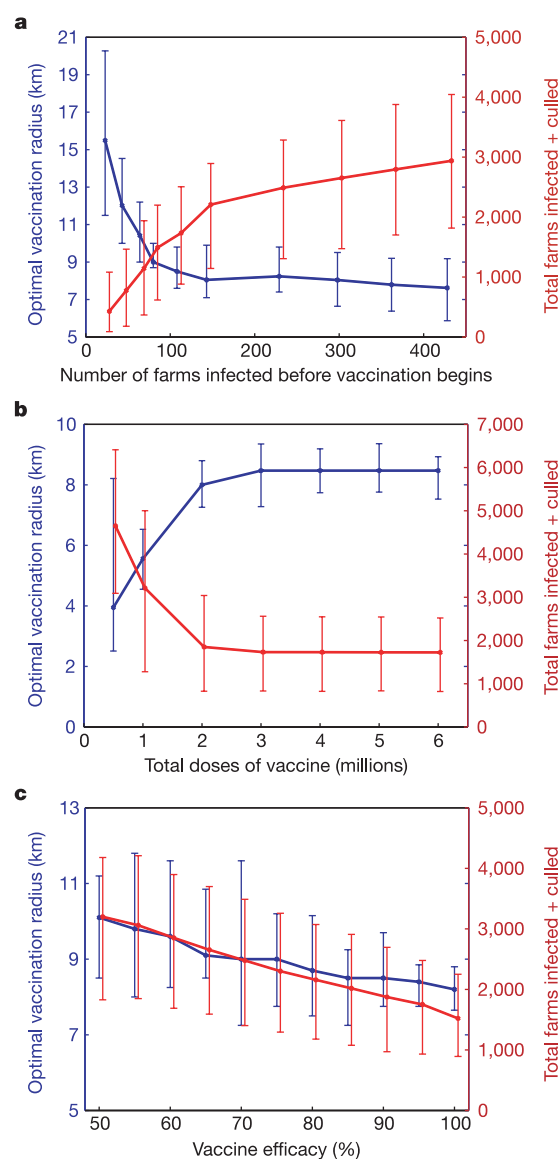


Figure 2 | The optimal vaccination radius and the associated epidemic impact. Results were determined (together with 95% confidence intervals) using an outside-in vaccination strategy and vaccination capacity of 35,000 cattle per day. **a**, The number of farms infected before vaccination begins is varied—capturing early or late detection and start of vaccination. **b**, The total number of doses of vaccine available is varied. **c**, The efficacy of the vaccine is varied. R_{inner} is zero in all simulations. Each point represents the results of 40,000 epidemic simulations. Similar results for the shortest-distance prioritization are given in the Supplementary Information.

Table 1 | Effect of strain variation on optimal vaccination radius

Risk factor	Optimal ring size (km)	
	Halved	Doubled
Sheep transmissibility	8.5 (7.3–0.8)	7.5 (6.8–8.1)
Cattle transmissibility	7.0 (6.0–8.5)	7.0 (6.3–7.5)
Sheep susceptibility	7.5 (6.5–8.4)	8.0 (6.5–10.0)
Cattle susceptibility	8.0 (6.3–11.0)	6.5 (5.8–7.6)
Kernel width (keeping total transmission constant)	6.0 (4.5–7.5)	8.0 (7.0–9.3)

The effect of varying epidemiological parameters, mimicking variation due to alternative strains, upon the optimal position of the outer vaccination ring. 95% confidence intervals are shown in parentheses.

other logistical and biological factors on the optimal vaccination radius and corresponding epidemic impact (Fig. 2). Figure 2a measures the delay before the disease is first detected combined with the delay before the onset of vaccination in terms of the average number of farms infected before vaccination begins. Prompt detection of the epidemic and a rapid decision to vaccinate allows larger vaccination rings to be implemented around each IP and substantially reduces the epidemic size, echoing more general principles^{1,12}. Any supply of vaccine will be limited because of the impracticality of storing large amounts of vaccine for each virus type; also, vaccine efficacy will vary between strains and may be considerably reduced if there is a discrepancy between the invading strain and the closest available vaccine. Both of these have implications for the success of any vaccination campaign^{13–15}. When the supply of vaccine is below three million doses, the vaccine ring must be reduced in size to prevent premature exhaustion of this supply, which in turn increases the epidemic impact (Fig. 2b). In contrast, although a reduction in vaccine efficacy (which may act as a surrogate for the farm-level effects of vaccination) also leads to a greater epidemic impact, it has a more limited effect on the optimal ring size (Fig. 2c).

The epidemiological characteristics within the model have been parameterized using the data from the 2001 epidemic. However, should foot-and-mouth disease enter the UK in future, it will not necessarily be the same strain and so its epidemiological characteristics could be substantially different. We therefore calculate the optimal vaccination ring (for the simple outside-in prioritization) while varying the susceptibility and transmissibility of cattle and sheep, and the width of the dispersal kernel, by a factor of two (Table 1). Although the average epidemic size is highly dependent upon these epidemiological parameters, the effect on the optimal radius, R_{outer} , is very small. We stress that these results are illustrative, because the true variation in strain characteristics has yet to be quantified, but do reflect the general robustness of the optimal ring size. We note, however, that strains that are effectively transmitted as air-borne plumes may give rise to dramatic changes to the

transmission kernel, and thus could invalidate the use of localized vaccination.

Ring vaccination is a form of targeted control that uses spatial proximity as a risk factor, but other refinements to the targeting could be used^{16,17}. Table 2 shows that whether vaccination is prioritized outside-in or inside-out within each day has very little effect on the optimal ring size or the associated epidemic impact. Surprisingly, if all prioritization is completely ignored and farms within a given radius are vaccinated at random there is a slight improvement over outside-in. This is because local herd-immunity (the threshold population level of immunity needed to prevent disease spread) around each IP is reached more quickly as it does not require the vaccination of all cattle. Identifying farms within a given radius, but prioritizing vaccination towards those farms with the largest numbers of cattle or livestock (rather than prioritizing in chronological order) again increases the optimal ring size and decreases the epidemic impact. This improvement mirrors the general epidemiological tenet that targeting of control measures can provide a substantial increase in effectiveness.

A variety of previous models have shown the success of CP culling in the absence of vaccination^{1,4,5} and this is confirmed by our simulations. However, for all vaccination priorities investigated here the inclusion of additional CP culling leads to the total loss of more farms (though not necessarily more livestock). We postulate that this loss of efficiency is because a high proportion of farms that are CP culled are (at least partially) protected by vaccination, and therefore the effect of CP culling on the spread of the infection is diminished compared to a situation with no vaccination. This means that the reduction in cases owing to CP culling is smaller than the number of additional culls performed, and we therefore conclude that CP culling is not advantageous when combined with vaccination.

We next turned to new strategies for optimizing the balance between managing existing and new foci of infection. A simple and very successful strategy is that of prioritizing farms for vaccination purely by their proximity to any previously infected premises, while vaccinating at capacity every day. The resultant epidemic impact is smaller than for all other prioritizations investigated (Table 2). This strategy has two main advantages. Firstly, an optimal ring size does not have to be determined, because this strategy utilises the full logistical resources that are available on any day. Therefore any change in logistical constraints does not necessitate a change in the vaccination policy. Secondly, vaccination in the immediate vicinity of a new IP is rapid, thus focusing control on those farms that are at the most immediate risk. The benefits gained from this targeting approach are not particularly sensitive to the precise ordering of vaccination, the number of cattle that can be vaccinated per day or the initial seeding of the epidemic. A further improvement is achieved if the prioritization is in terms of the shortest distance to

Table 2 | Effect of control strategies on optimal vaccination radius

Vaccination priority	Optimal ring size (km)	IPs + DCs + CPs	Total animals culled (millions)
Earliest IP reporting date (outside-in)	8.5 (7.8–9.6)	1,752 (928–2544)	2.07 (1.04–3.23)
Earliest IP reporting date (inside-out)	8.6 (8.0–0.5)	1,791 (994–2469)	2.10 (1.04–3.30)
Random order	9.0 (7.8–11.2)	1,688 (883–2475)	2.02 (0.94–3.13)
Largest cattle farms	10.5 (8.8–11.9)	1,535 (745–2339)	1.74 (0.71–2.85)
Largest mixed farms	12.5 (10.3–15.4)	1,343 (696–2025)	1.42 (0.61–2.28)
Shortest distance to any IP	NA	1,254 (664–1954)	1.50 (0.67–2.46)
Shortest distance to any IP or DC	NA	1,192 (626–1860)	1.42 (0.65–2.35)
Shortest distance to any IP or DC identified within previous ten days	NA	1,088 (582–1726)	1.28 (0.60–2.17)
No vaccination (+CP culling)	NA	4,220 (2149–6431)	3.58 (2.32–5.02)
Earliest IP reporting date (+CP culling)	8.0 (6.6–9.8)	1,955 (1176–2803)	1.65 (0.80–2.54)
Largest mixed farms (+CP culling)	11.3 (9.7–22.5)	1,607 (1033–2277)	1.27 (0.72–1.93)
Shortest distance to IP (+CP culling)	NA	1,528 (895–2272)	1.69 (0.73–2.70)

The effect of vaccination priority and additional control strategies upon the optimal position of the outer vaccination ring together with the associated epidemic impact (IPs, and also DCs and CPs if appropriate) and the total number of animals culled. 95% confidence intervals are shown in parentheses. NA, not applicable.

any IP or DC identified within the past ten days, as this also targets vaccination around DCs that are suspected of being infected and ignores regions of the country that no longer pose any risk.

We have shown that reactive vaccination could be a very powerful tool for combating FMD, and potentially other locally transmitted pathogens, when combined with efficient IP and DC culling and animal movement restrictions. As ever, rapid implementation is essential. A policy that prioritizes vaccination in terms of the shortest distance to any reported infection or DC has strong epidemiological benefits over traditional fixed-radius policies, especially in terms of balancing local and regional control and making full use of the daily vaccination capacity. It would also be easy to achieve practically, because optimization for a particular set of logistical constraints is unnecessary. This specific result echoes a more general trend, that detailed data-driven models are becoming an increasingly important tool in determining optimal epidemic control policies.

METHODS

The Model. The probability that a susceptible farm i is infected on a given day, D , is given by the equation:

$$P_{i,D} = 1 - \exp \left(-S^{(i)} \sum_{j \in \text{Infectious}(D)} T^{(j)} K(d_{ij}) \right) \quad (1)$$

$$S^{(i)} = s_{\text{cow}} n_{\text{cow}}^{(i)} + s_{\text{sheep}} n_{\text{sheep}}^{(i)} \quad T^{(j)} = t_{\text{cow}} n_{\text{cow}}^{(j)} + t_{\text{sheep}} n_{\text{sheep}}^{(j)}$$

where $S^{(i)}$ and $T^{(j)}$ refer to the susceptibility and transmissibility of uninfected farm i and infectious farm j , respectively. These farm-level measures are in turn calculated from species-level parameters where s_A and t_A refer to the susceptibility and transmissibility of species A , and $n_A^{(i)}$ is the number of species A on farm i . These animal-level parameters are estimated from the 2001 data by matching to the cumulative number of cases and culls in five distinct regions (Cumbria, Devon, Wales, Scotland and the rest of England). The best-fit parameters for Cumbria exemplify the general pattern: $t_{\text{cow}} = 7.7 \times 10^{-7}$, $t_{\text{sheep}} = 5.1 \times 10^{-7}$, $s_{\text{cow}} = 10.5$, $s_{\text{sheep}} = 1$ (see Supplementary Information for full parameters and confidence intervals). Pigs have been ignored in this model for simplicity because there is insufficient data from the 2001 epidemic to determine precise parameter values or even to show that the parameters are significantly different from zero. K is the infection kernel and d_{ij} is the distance between susceptible farm i and infectious farm j . The kernel is calculated from the contact tracing performed by DEFRA during 2001 after movement restrictions, which provides a distribution of distances between infecting farms². We note that at present in the UK, only the residential position of the owner of the farm is recorded, which could be separate from the farm itself and also distant from where livestock is kept. However, large discrepancies in position affect only a small fraction of the total data and sensitivity analysis shows that the general results are unaffected by small errors in the positional data.

This epidemiological framework is coupled with the prompt culling of livestock on IPs (within 24 hours) and on DCs (within 48 hours). The model therefore mimics the rapid speed and significant levels of local control that were attained during the latter part of the 2001 outbreak, representing what it is hoped could be achieved from the detection of an epidemic in the future. The simulations are initialized with the exact conditions of 23 February 2001 (including infected, infectious and reported farms); in any future outbreak it is hoped that the disease will be detected earlier so that the spread is less disseminated (Fig. 2a).

Calculation of dangerous contacts and contiguous premises. In practice, DCs are identified for each IP on a case-by-case basis, and are based on veterinarian judgement of risk factors and known activities, such as the movement of vehicles. In our model, DCs are determined stochastically, accounting for infection risk factors and known transmission events (see Supplementary Information) and parameterized to match the best sustainable levels achieved during 2001.

CPs are determined by a comprehensive knowledge of the farm geography and are defined as farms that share a common boundary—in practice, this is again determined on a case-by-case basis using local maps and knowledge. However, the data on farms within the UK only identifies the location of the owner of the farm. We approximate CPs by tessellating around each point location, taking into account the known area of each farm, to obtain a surrogate set of adjacent farms. Clearly this set of farms will not necessarily be the set of true CPs^{1,2}, but this approximation will capture many of the elements of local proximity.

Vaccination assumptions. There are a set of standard assumptions that are made

throughout the paper about the characteristics of vaccination. First, we assume that only cattle are vaccinated—vaccination of other species, such as sheep, would substantially limit the number of farms that could be vaccinated per day and therefore limit the effect of vaccination (see Supplementary Information). Unless otherwise stated, the vaccine efficacy is taken to be 90%, and a four-day delay from vaccination to immunity is assumed. Therefore, after four days, 90% of cattle on a given farm become totally immune and the remainder are totally susceptible. We make the pessimistic assumption that during the four days between vaccination and immunity, all cattle are completely susceptible and if infected, the disease progresses in the same way as for non-vaccinated cattle. A farm in which 90% of the cattle are immune has the same transmission and susceptibility properties as a farm with 10% of the number of cattle. In general, we assume that 35,000 animals can be vaccinated per day (unless otherwise stated) and there is no limit on the number of farms that can be vaccinated per day. This assumption results in vaccination of 260 farms per day (on average) for shortest-distance ring vaccination. The latest DEFRA estimates suggest that a maximum of 300 farms could be vaccinated per day, so our limit of 35,000 animals per day is reasonable. Vaccination is assumed to begin five days after the disease is first detected, in agreement with the DEFRA contingency plan⁶.

Received 16 June; accepted 29 September 2005.

1. Keeling, M. J. *et al.* Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813–817 (2001).
2. Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G. & Grenfell, B. T. Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142 (2003).
3. Ferguson, N. M., Donnelly, C. S. & Anderson, R. M. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* **292**, 1155–1160 (2001).
4. Ferguson, N. M., Donnelly, C. S. & Anderson, R. M. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542–548 (2001).
5. Morris, R. S., Wilesmith, J. W., Stern, M. W., Sanson, R. L. & Stevenson, M. A. Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in Great Britain, 2001. *Vet Record* **149**, 137–144 (2001).
6. DEFRA Foot and Mouth Disease Contingency Plan (<http://www.defra.gov.uk/footandmouth/contingency/index.htm>) (2004).
7. Anderson, I. *Foot and Mouth Disease 2001: Lessons to be Learned Inquiry* (The Stationery Office, London, 2002).
8. Council Directive 2003/85/EC. *Official J. Eur. Union* **L306**, 46, 1–87 (2003).
9. Muller, J., Schonfisch, B. & Kirkilionis, M. Ring vaccination. *Math. Biol.* **41**, 143–171 (2000).
10. Barnett, P. V. & Carabin, H. A review of emergency foot-and-mouth disease (FMD) vaccines. *Vaccine* **20**, 1505–1514 (2002).
11. Woolhouse, M. E. J., Haydon, D. T., Pearson, A. & Kitching, R. P. Failure of vaccination to prevent outbreaks of foot-and-mouth disease. *Epidem. Infect.* **116**, 363–371 (1996).
12. Ferguson, N. M. *et al.* Planning for smallpox outbreaks. *Nature* **425**, 681–685 (2003).
13. McLean, A. R. & Blower, S. M. Modelling HIV vaccination. *Trends Microbiol.* **3**, 458–463 (1995).
14. Williams, J. R., Nokes, D. J. & Anderson, R. M. Targeted hepatitis B vaccination—a cost effective immunisation strategy for the UK? *J. Epidemiol. Comm. Health* **50**, 667–673 (1996).
15. Woolhouse, M. E. J., Haydon, D. T. & Bundy, D. A. P. The design of veterinary vaccination programmes. *Vet. J.* **153**, 41–47 (1997).
16. Woolhouse, M. E. J. *et al.* Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl Acad. Sci. USA* **94**, 338–342 (1997).
17. Matthews, L. *et al.* Neighbourhood control policies and the spread of infectious disease. *Proc. R. Soc. Lond. Ser. B* **270**, 1659–1666 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by the Wellcome Trust.

Author Contributions M.J.T. and M.J.K. were responsible for the model formulation and analysis of results; N.J.S. provided helpful discussions throughout; D.J.S. generated cleaned demographic and epidemic data; R.D. and S.P.B. provided vital statistical input; M.E.J.W. and B.T.G. were instrumental in the initial development of the project. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.J.K. (m.j.keeling@warwick.ac.uk).

Sexual reproduction selects for robustness and negative epistasis in artificial gene networks

Ricardo B. R. Azevedo¹, Rolf Lohaus¹, Suraj Srinivasan¹, Kristen K. Dang² & Christina L. Burch³

The mutational deterministic hypothesis for the origin and maintenance of sexual reproduction posits that sex enhances the ability of natural selection to purge deleterious mutations after recombination brings them together into single genomes¹. This explanation requires negative epistasis, a type of genetic interaction where mutations are more harmful in combination than expected from their separate effects. The conceptual appeal of the mutational deterministic hypothesis has been offset by our inability to identify the mechanistic and evolutionary bases of negative epistasis. Here we show that negative epistasis can evolve as a consequence of sexual reproduction itself. Using an artificial gene network model^{2,3}, we find that recombination between gene networks imposes selection for genetic robustness, and that negative epistasis evolves as a by-product of this selection. Our results suggest that sexual reproduction selects for conditions that favour its own maintenance, a case of evolution forging its own path.

A century of genetic research has revealed two general properties of spontaneous mutations with detectable effects on fitness: most of them are deleterious, and they frequently interact with each other^{4,5}. Many types of interactions are possible, including directional epistasis, in which the average effect of spontaneous mutations changes in the presence of other mutations in the genome⁶. Directional epistasis can be either negative (synergistic) or positive (antagonistic), depending on whether the average effect of mutations becomes more or less harmful, respectively, as the number of other mutations in the genome increases (Fig. 1). Directional epistasis holds particular interest for evolutionary biologists because it is expected to determine the outcome of multiple evolutionary processes, notably the evolution of sex and recombination¹. Empirical studies on a variety of organisms have reported every conceivable form of directional epistasis: negative^{7–9}, positive^{6,10} and no significant directional epistasis^{11,12}. These mixed results have not helped to clarify either the mechanistic or evolutionary causes of directional epistasis¹³.

In contrast, evolutionary simulations using computational models of RNA secondary structure¹⁴, viral replication¹⁵ and artificial life¹⁴ have demonstrated that the average strength and direction of epistasis can be shaped by natural selection. One mechanism by which epistasis evolves in these models¹³ is through a negative correlation among genotypes between the extent of genetic robustness (or genetic canalization, measured as the insensitivity of a phenotype to mutation) and the direction of epistasis. As a consequence, selection for higher robustness produces a correlated response in the strength of epistasis in all three models, towards either weaker positive or stronger negative epistasis^{14,15}. The repeatability of this result in models of different biological systems suggests that the strength and direction of epistasis observed in living organisms depend on their history of selection for genetic robustness.

Theory predicts that traits can evolve to be robust to genetic perturbations (that is, mutation and recombination) under a variety

of selective regimes^{16–18}, as long as the following two conditions are met: genes must interact to determine the trait^{17–19}, and the population must contain sufficient genetic variation¹⁸. Whereas the former condition is inherent to particular organisms, the latter condition will depend on population genetic parameters such as the mutation and recombination rates. Experimental tests of these predictions using computational models confirm that high mutation rates, such as those experienced by RNA viruses, favour the evolution of genetic robustness^{2,3,18,20}. Sexual reproduction (that is, increased recombination) is also expected to impose stronger selection for genetic robustness than asexual reproduction^{21,22}, but this hypothesis has never been tested experimentally²¹.

To test this hypothesis, and to determine whether the evolution of genetic robustness is accompanied by the evolution of negative epistasis, we return to the computational model of genetic networks used in two previous studies^{2,3}. We chose this model primarily because it explicitly incorporates one of the key characteristics required for the evolution of robustness^{17–19}—genetic interactions. Furthermore, empirical data from biological systems has consistently suggested that extant gene networks are robust to changes in biochemical rate parameters and levels of gene activity^{19,23}. Previous work with this model has shown that genetic robustness (again, measured as robustness to mutation) evolves readily if networks are subjected to selection for the production of a stable gene expression pattern^{2,3}. Here we explore the extent to which recombination contributed to the evolution of genetic robustness in this model, and ask whether recombination, through its effect on robustness^{2,21,22}, can cause the direction of epistasis to evolve.

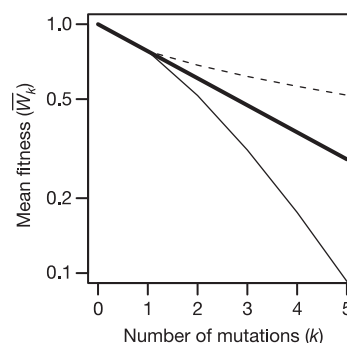


Figure 1 | Types of directional epistasis for deleterious mutations. Three hypothetical relationships between fitness (log scale) and number of deleterious mutations are plotted. All relationships depicted have the same mutational robustness ($\bar{W}_1 = 0.78$) but different directions of epistasis: negative epistasis (plain line, concave downwards; $1 - \beta < 0$), no directional epistasis (bold, straight line; $1 - \beta = 0$) and positive epistasis (dashed line, concave upwards; $1 - \beta > 0$).

¹Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, USA. ²Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7575, USA. ³Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3280, USA.

Briefly, the model^{2,3} represents individuals as networks of N interacting transcriptional regulators (Fig. 2a). The genotype of an individual is represented by an $N \times N$ matrix R , whose elements r_{ij} describe the regulatory effect of the product of gene j on the expression of gene i (Fig. 2b). The number of regulatory interactions is determined by a connectivity parameter (c) that specifies the proportion of non-zero matrix elements. This matrix of regulatory relationships acts on gene expression patterns, which are represented by a state vector $S(t)$, whose elements $s_i(t)$ describe the expression states of genes $i = 1, 2, \dots, N$ at time t . The expression state of a gene can vary continuously between complete repression, $s_i(t) = -1$, and complete activation, $s_i(t) = 1$. Gene expression states change over time according to the following equation:

$$s_i(t+1) = f \left[\sum_{j=1}^N r_{ij} s_j(t) \right] \quad (1)$$

where $f(x)$ is a sigmoidal filter function² that determines how the total regulatory input influences gene expression (see Supplementary Fig. 1a). Development is modelled as the progression from an initial gene expression state to an equilibrium gene expression pattern (Fig. 2c; see Methods). In this model, genotypes that achieve any stable, fixed-point equilibrium expression pattern are considered developmentally stable², and therefore viable. Genotypes that do not achieve a stable equilibrium (for example, oscillatory gene expression) are considered inviable².

The mechanistic underpinnings of this model allow *a priori* predictions about the evolution of genetic robustness. In this model, the reproductive success of a viable genotype (that is, its fitness) is the proportion of its offspring that are also viable. When offspring are produced asexually, and differ from their parents only by mutations, the fitness of a genotype is given by:

$$W_{\text{asex}} = \sum_{k=0}^{cN^2} \phi_k \bar{W}_k \quad (2)$$

where $\phi_k = \mu^k e^{-\mu} / k!$ is the Poisson probability that offspring

acquire k mutations when the mutation rate per individual network per generation is μ , and $\bar{W}_k$ is the mean fitness of the genotype after the addition of k mutations. For mutation rates $\mu \leq 0.1$, terms for $k > 1$ can be effectively ignored, so that $W_{\text{asex}} \approx 1 - \mu + \mu \bar{W}_1$. $\bar{W}_1$ is our measure of mutational robustness, the probability that a genotype with one mutation is viable. Therefore, given a sufficiently high mutation rate, asexually reproducing networks are expected to evolve mutational robustness. In contrast, when offspring are produced via sexual reproduction, and differ from their parents primarily as a result of recombination, the fitness of a genotype is given by $W_{\text{sex}} = W_{\text{asex}}(1 - L)$, where L is the probability that a mating with a random individual in the population will result in an inviable offspring, a measure of the recombination load²⁴. Therefore, sexual populations should experience selection for two distinct types of genetic robustness: mutational robustness and recombinational robustness.

In order to explore the effect of sexual reproduction on the evolution of genetic robustness, we first investigated the behaviour of this model using conditions that are known to produce robustness to mutation² ($\mu = 0.1$, $N = 10$ genes, $c = 0.75$; see Methods), varying only the reproductive mode from sexual to asexual. Fifty clonal populations of 500 individuals were founded by different randomly generated, viable genotypes. Each population was subjected to selection for the ability to produce a stable gene expression pattern and allowed to evolve separately via sexual and asexual reproduction. We monitored evolution until an equilibrium level of mutational robustness was achieved. Contrary to earlier claims^{2,3}, our simulations show that sexual reproduction has a substantive effect on the evolution of mutational robustness (Fig. 3a). Although mutational robustness increased in asexual populations, it reached a significantly lower equilibrium value than in sexual populations (paired t -test: $t = 31.0$, 49 degrees of freedom (d.f.), $P < 0.0001$). An investigation of epistasis in these evolved populations revealed that sexual reproduction also had a qualitative effect on the evolution of directional epistasis ($t = 23.6$, 49 d.f., $P < 0.0001$). The magnitude of epistasis evolved regardless of reproductive mode, but the direction of epistasis only changed when reproduction was sexual. At equilibrium, asexual populations exhibited average positive epistasis of a reduced magnitude, whereas sexual populations exhibited negative epistasis.

Why does sexual reproduction cause the evolution of increased mutational robustness? Sexual reproduction is not expected to increase the strength of selection for mutational robustness directly. However, it is expected to select for recombinational robustness, and this could cause a correlated response in mutational robustness. We devised two experiments to test this hypothesis. In the first experiment, we investigated whether the effect of sexual reproduction on mutational robustness depended on the high mutation rate ($\mu = 0.1$). Theory predicts that mutational robustness will evolve through the direct action of selection only if μ is greater than the reciprocal of the effective population size⁷ (that is, $\mu > 0.002$ in our simulations). Thus, we tested our hypothesis by re-running the simulations at a mutation rate of $\mu = 0.002$ (Fig. 3a). At this low mutation rate, mutational robustness failed to evolve in asexual populations within 50,000 generations. However, mutational robustness did increase significantly in sexual populations within 20,000 generations. The inability of asexual populations to respond to selection for mutational robustness confirms that selection acting directly on mutational robustness is ineffective when $\mu = 0.002$. Thus, the mutational robustness that evolved in these sexual populations did not evolve through the direct action of selection. Rather, it must have evolved as a correlated response to selection for recombinational robustness, the only other source of selection in these simulations.

In the second experiment we constructed genetically variable populations (see Supplementary Methods) and allowed them to evolve in the absence of new mutations ($\mu = 0$), that is, in the absence of selection for mutational robustness. In this experiment,

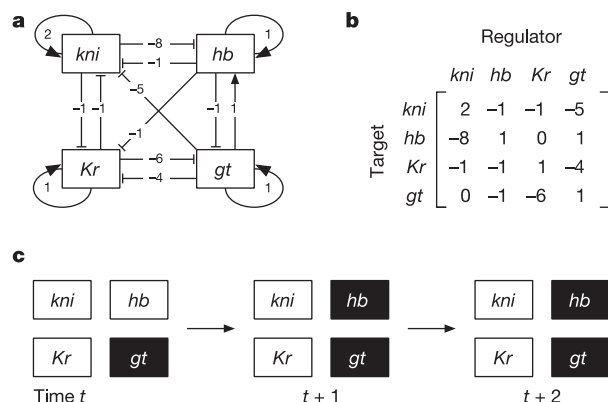


Figure 2 | Application of our network model to the gap gene system of *Drosophila melanogaster*. **a**, Network representation of the regulatory interactions between four gap genes²⁹ (*gt*, *giant*; *hb*, *hunchback*; *kni*, *knirps*; *Kr*, *Krüppel*). Activations and repressions are denoted by arrows and bars, respectively. Numbers indicate the relative interaction strengths³⁰. **b**, Interaction matrix (R) representing the network in **a**. The element in row i and column j (r_{ij}) denotes the regulatory effect of the product of gene j on the expression of gene i . **c**, Graphical representation of the gene expression states of each gap gene over three successive time steps. For the purpose of this illustration we consider gene i to be ON (filled box) if $s_i(t) > 0$, and OFF (open box) if $s_i(t) \leq 0$. The change in gene expression pattern matches events at ~80% anterior-posterior position in the *Drosophila* embryo between early and mid cleavage cycle 14A (ref. 29). Successive iterations beyond the $t + 1$ step do not change the gene expression pattern, the hallmark of a stable equilibrium.

sexual populations showed significant increases in mutational robustness, whereas asexual populations did not (Fig. 3b). These results confirm our hypothesis because sexual and asexual populations differed only by the presence and absence, respectively, of selection for recombinational robustness. By manipulating the amount of genetic variation present in the founder populations, we also showed that the evolutionary response in mutational robustness increased with the strength of selection for recombinational robustness (that is, with the magnitude of the recombination load, L ; Supplementary Fig. 2).

Directional epistasis evolved in both of these experiments (Fig. 3) in a similar manner to the initial simulations. Conditions that showed no change in mutational robustness also showed no change in directional epistasis. However, conditions that caused an evolutionary response in mutational robustness also caused the evolution of negative, or less positive, epistasis. Taken together, these results confirm that mutational robustness and negative epistasis both evolved in response to selection for recombinational robustness.

The most likely explanation for the evolution of negative epistasis in these simulations is that epistasis evolved as a correlated response to selection for genetic robustness. The direction of epistasis was negatively correlated with mutational robustness among a random sample of viable gene networks (Supplementary Fig. 3). Similar correlations were found in digital organisms and RNA secondary structure¹⁴, supporting the theoretical prediction^{14,25} that it is impossible to change genetic robustness and the direction of epistasis independently.

Although we recognize that our model describes a simplified view of transcriptional regulation, it captures an important feature of real genetic regulatory systems: genetic interactions are abundant, causing mutations to have different effects depending on the genetic background in which they arise. We propose that sexual reproduction

will favour the evolution of increased genetic robustness and, therefore, negative epistasis in any system with two key properties: numerous genetic interactions and abundant genetic variation—both known requirements for the evolution of genetic robustness^{17–19}. Consistent with this proposal, the parameters that determine the number of genetic interactions (connectivity and gene number) and the amount of genetic variation (mutation rate, population size and the strength of stabilizing selection) all influenced whether negative epistasis evolved in our simulations (Fig. 3; see also Supplementary Figs 4–7). In contrast, the network topology, the shape and variance of the mutational distribution, and environmental stochasticity did not qualitatively affect the outcome (Supplementary Figs 1, 5, 8 and 9). Most notably, negative epistasis failed to evolve in networks that were both small and sparsely connected (Supplementary Figs 5 and 6). However, the requirements for the evolution of negative epistasis were not too restrictive. Sexual reproduction produced negative epistasis even in small networks as long as they were sufficiently connected, and in sparsely connected networks as long as they contained a sufficient number of genes (Supplementary Figs 5 and 6). Stabilizing selection acting on the gene expression pattern also prevented the evolution of negative epistasis, but only when it was exceptionally strong (Supplementary Fig. 7). The wealth of genetic interactions in the transcriptional networks of real organisms²⁶ and the abundance of genetic variation in natural populations²⁷ suggest that negative epistasis will evolve in many sexually reproducing organisms.

The evolution of negative epistasis in our simulations is remarkable because it suggests that sexual reproduction selects for conditions that favour its own maintenance. Because negative epistasis enhances the ability of natural selection to purge deleterious mutations in sexual populations, our results could explain the maintenance of sexual reproduction in the face of its numerous

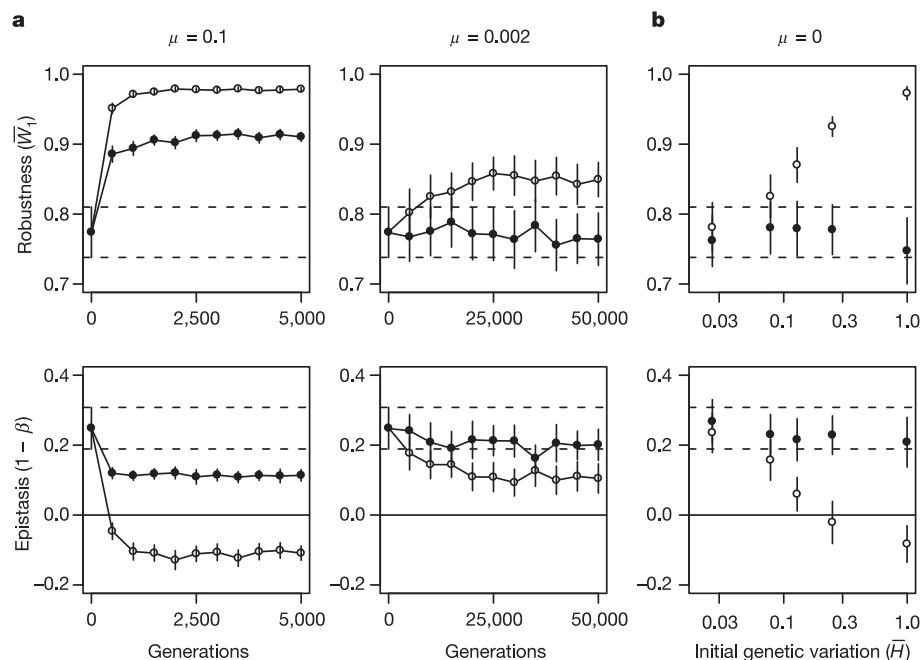


Figure 3 | Sexual reproduction selects for mutational robustness and negative epistasis. **a**, Selection for the ability to produce any stable gene expression pattern was imposed on 50 replicate populations subjected to high ($\mu = 0.1$) and low ($\mu = 0.002$) mutation rates. Plots show the average evolutionary responses in robustness to mutation and direction of epistasis. **b**, The 50 individuals used to found the homogeneous populations described in **a** were used to found new populations of 500 individuals with 1, 3, 5, 10 or 75 random mutations each. These populations were then allowed to evolve without the occurrence of new mutations ($\mu = 0$) until genetic variation was

exhausted. We plot the robustness to mutation and the direction of epistasis at equilibrium (that is, after evolution stops) against the initial genetic variation in each treatment (see Supplementary Methods). Each population in **a** and **b** was evolved under either asexual (filled circles) or sexual (open circles) reproduction. Data are expectations and 95% confidence intervals for the median value among the 500 individuals in each population. Dashed lines indicate the 95% confidence intervals for the mean of the 50 founder networks.

costs²⁸. We hypothesize that sexual reproduction enabled evolution of the robustness apparent in the developmental networks of multicellular organisms and that negative epistasis should be associated with robustness in these systems. If these hypotheses are correct, they will help to explain the prevalence of sexual reproduction among living organisms.

METHODS

Network model. Networks are generated by randomly filling the entries of the R matrix (for example, Fig. 2b) with $(1 - c)N^2$ zeros and cN^2 standard normal random variates. A corresponding initial gene expression pattern, $S(0)$, is created for each network by randomly setting each $s_i(0)$ to either -1 or 1 . Development begins with the initial gene expression pattern, $S(0)$, and proceeds through 100 iterations of equation (1). We determined that an equilibrium steady state was achieved when the following criterion was met²:

$$\sum_{\theta=-10}^t D[S(\theta), \bar{S}(t)] \leq 10^{-3}, \quad \text{where } D[S, S'] = \frac{1}{4N} \sum_{i=1}^N (s_i - s'_i)^2$$

is a measure of the difference between the gene expression patterns S and S' , and $\bar{S}(t)$ is the average of the gene expression levels over the time interval from $t - 10$ to t . The ability of a genotype to reach equilibrium within 100 iterations is termed developmental stability².

Evolution. In a typical evolutionary simulation, a single random individual capable of producing a stable gene expression pattern is cloned to generate a population of 500 identical individuals. In an asexually reproducing population, offspring are generated by picking an individual at random from the population and allowing it to produce a clone of itself, such that each non-zero entry in the R interaction matrix mutates (replacement with an independent standard normal random variate) with probability $\mu/(cN^2)$. In our model, mutations should be viewed as acting on the cN^2 *cis*-regulatory elements, not the coding sequences of the N genes themselves; in addition, mutations cannot alter the number of genes, or establish new interactions between genes. Only offspring capable of producing a stable gene expression pattern survive. This process is repeated until 500 developmentally stable individuals are produced, which go on to found the following generation. In a sexual population, offspring are generated by picking two individuals at random from the population, and selecting rows of the R matrices from each parent with equal probability (analogous to free recombination between units formed by each gene and its *cis*-regulatory elements, but with no recombination within regulatory regions), while allowing each non-zero entry to mutate as above. Each selective regime was applied to a fixed panel of 50 replicate populations, each derived from a single independently generated random individual and initial gene expression pattern; simulations were run for as long as was necessary to obtain an equilibrium (that is, no significant change) in the second half of the simulation. In each simulation, all individuals experience the same initial gene expression pattern as the founder individual.

Robustness and epistasis. The mean effects of k mutations on fitness were modelled by the relationship^{14,15}: $\log(\bar{W}_k) = -\alpha k^\beta$ (Fig. 1, equation (2)). Mutational robustness and directional epistasis were measured by $\bar{W}_1$ and $1 - \beta$, respectively. To estimate these parameters for a given genotype, we generated 100 individuals with five successive rounds of random mutations each, and measured the proportion of viable genotypes, $\bar{W}_k$, with $k = 1, 2, \dots, 5$ mutations. We modelled $\bar{W}_k$ using a generalized linear model with complementary log-log link and a binomial error structure²⁹:

$$\log[-\log(\bar{W}_k)] = \log(\alpha) + \beta \log(k).$$

$\bar{W}_1$ was measured directly and β was estimated using maximum likelihood.

Received 1 October; accepted 22 November 2005.

1. Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).
2. Siegal, M. L. & Bergman, A. Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl Acad. Sci. USA* **99**, 10528–10532 (2002).
3. Wagner, A. Does evolutionary plasticity evolve? *Evolution* **50**, 1008–1023 (1996).

4. Lynch, M. *et al.* Perspective: Spontaneous deleterious mutation. *Evolution* **53**, 645–663 (1999).
5. Whitlock, M. C., Phillips, P. C., Moore, F. B. G. & Tonsor, S. J. Multiple fitness peaks and epistasis. *Annu. Rev. Ecol. Syst.* **26**, 601–629 (1995).
6. Burch, C. L. & Chao, L. Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics* **167**, 559–567 (2004).
7. de Visser, J. A. G. M., Hoekstra, R. F. & van den Ende, H. An experimental test for synergistic epistasis and its application in *Chlamydomonas*. *Genetics* **145**, 815–819 (1997).
8. Mukai, T. The genetic structure of natural populations of *Drosophila melanogaster*. VII. Synergistic interaction of spontaneous mutant polygenes controlling viability. *Genetics* **61**, 749–761 (1969).
9. Whitlock, M. C. & Bourguet, D. Factors affecting the genetic load in *Drosophila*: synergistic epistasis and correlations among fitness components. *Evolution* **54**, 1654–1660 (2000).
10. Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M. & Petropoulos, C. J. Evidence for positive epistasis in HIV-1. *Science* **306**, 1547–1550 (2004).
11. de Visser, J. A. G. M., Hoekstra, R. F. & van den Ende, H. Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. *Evolution* **51**, 1499–1505 (1997).
12. Elena, S. F. & Lenski, R. E. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
13. Michalakis, Y. & Roze, D. Evolution. Epistasis in RNA viruses. *Science* **306**, 1492–1493 (2004).
14. Wilke, C. O. & Adami, C. Interaction between directional epistasis and average mutational effects. *Proc. R. Soc. Lond. B* **268**, 1469–1474 (2001).
15. You, L. & Yin, J. Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage T7. *Genetics* **160**, 1273–1281 (2002).
16. Kawecki, T. J. The evolution of genetic canalization under fluctuating selection. *Evolution* **54**, 1–12 (2000).
17. Rice, S. H. The evolution of canalization and the breaking of von Baer's laws: Modeling the evolution of development with epistasis. *Evolution* **52**, 647–656 (1998).
18. Wagner, G. P., Booth, G. & Bagheri-Chaichian, H. A population genetic theory of canalization. *Evolution* **51**, 329–347 (1997).
19. Nijhout, H. F. The nature of robustness in development. *Bioessays* **24**, 553–563 (2002).
20. Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. & Adami, C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* **412**, 331–333 (2001).
21. de Visser, J. A. G. M. *et al.* Perspective: Evolution and detection of genetic robustness. *Evolution* **57**, 1959–1972 (2003).
22. Stearns, S. C. The evolutionary links between fixed and variable traits. *Acta Paleontol. Pol.* **38**, 215–232 (1994).
23. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. III & Doyle, J. Robustness of cellular functions. *Cell* **118**, 675–685 (2004).
24. Charlesworth, B. & Barton, N. Recombination load associated with selection for increased recombination. *Genet. Res.* **67**, 27–41 (1996).
25. Wagner, G. P., Laubichler, M. D. & Bagheri-Chaichian, H. Genetic measurement of theory of epistatic effects. *Genetica* **102–103**, 569–580 (1998).
26. Milo, R. *et al.* Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
27. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3**, 11–21 (2002).
28. Maynard Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, Cambridge, 1978).
29. Crawley, M. J. *Statistical Computing* (Wiley, Chichester, 2002).
30. Jaeger, J. *et al.* Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. *Genetics* **167**, 1721–1737 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Flatt, Y. Fofanov, F. Galis, J. Kingsolver, A. Monteiro, M. Travisano and G. Wagner for discussions. The UH, UNC and NIH (grant to C.L.B.) provided financial support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.B.R.A. (razevedo@uh.edu).

Dissociable neural mechanisms supporting visual short-term memory for objects

Yaoda Xu¹ & Marvin M. Chun¹

Using visual information to guide behaviour requires storage in a temporary buffer, known as visual short-term memory (VSTM)¹, that sustains attended information across saccades and other visual interruptions. There is growing debate on whether VSTM capacity is limited to a fixed number of objects^{2,3} or whether it is variable^{4,5}. Here we report four experiments using functional magnetic resonance imaging that resolve this controversy by dissociating the representation capacities of the parietal and occipital cortices. Whereas representations in the inferior intraparietal sulcus (IPS) are fixed to about four objects at different spatial locations regardless of object complexity, those in the superior IPS and the lateral occipital complex are variable, tracking the number of objects held in VSTM, and representing fewer than four objects as their complexity increases. These neural response patterns were observed during both VSTM encoding and maintenance. Thus, multiple systems act together to support VSTM: whereas the inferior IPS maintains spatial attention over a fixed number of objects at different spatial locations, the superior IPS and the lateral occipital complex encode and maintain a variable subset of the attended objects, depending on their complexity. VSTM capacity is therefore determined both by a fixed number of objects and by object complexity.

Behavioural studies have reported that VSTM can store up to about four objects independent of the number of features that must be encoded for each object³, indicating that VSTM capacity might be limited to a fixed number of objects³. However, new evidence shows that fewer than four objects are held in VSTM when object complexity increases^{4,5}, indicating that VSTM capacity is variable and modulated by the complexity of visual objects encoded⁵.

A network of brain areas supports VSTM (that is, visual working memory), including the frontal/prefrontal, parietal and occipital cortices^{6–19}. Although the frontal/prefrontal cortex shows increased activity with memory load⁹, a recent study indicates that the (superior) IPS might have the most important function in determining VSTM capacity for visual objects^{17,18}, because functional magnetic resonance imaging (fMRI) activations within this region correlated most strongly with the number of objects held in VSTM (see also a related electrophysiological study¹⁹). However, these recent findings do not address the fundamental debate over whether neural processing capacity is limited by a fixed number of objects or whether it varies with increasing visual object complexity^{2–5}. In addition, although the role of the frontal/prefrontal cortices in the control and maintenance of VSTM has been well described^{7–14}, the nature of VSTM object representation in occipital and other parietal regions is less well understood and deserves further investigation^{6,11,15,16}.

The occipital visual areas have been frequently reported to be involved in VSTM tasks^{9–13,16}, revealing that VSTM storage may be mediated by the same brain structures processing perceptual

information^{15,16}. Among the different visual areas, the lateral occipital complex (LOC) is important in visual object perception such that it exhibits higher activations for objects than for noise images^{20–22} and participates in successful visual object recognition²³. Whether LOC activity directly correlates with the number of objects successfully retained in VSTM, however, has not been thoroughly established.

Besides LOC, a higher activation for objects than for noise images has also been reported in the inferior IPS, which has been attributed to greater visual attention to objects than to noise images²². Although previous research has indicated that parietal attention mechanisms are involved in VSTM for spatial information^{12,24}, given that attention is essential in VSTM for objects as well, the role of the parietal attention mechanisms in maintaining visual objects requires more detailed investigation.

To address these issues, using fMRI, we examined the roles of the parietal and occipital cortices in VSTM for visual objects. In the first experiment, observers were required to remember either simple or relatively complex (but still easily discriminable) shape features from the same set of objects. Observers viewed one to four, or six black, elongated objects in a sample display and, after a brief delay, detected a shape feature change in a test display. The simple and the relatively complex (hereafter referred to as ‘complex’) shape features were defined as the presence or absence of the hole in the centre of the objects and object outlines, respectively (Fig. 1a).

To evaluate behavioural responses, we applied Cowan’s *K* formula²⁵ and transformed change detection accuracies to VSTM capacity estimates as a function of display set size. To evaluate fMRI responses, we used a region of interest (ROI) approach²² and extracted averaged fMRI responses from three functionally defined ROIs (Fig. 2a). The superior IPS ROI was defined as voxels whose activations correlated with VSTM capacity in a separate colour VSTM task (Supplementary Fig. 1a) and whose Talairach coordinates²⁶ matched those reported previously¹⁷. LOC and inferior IPS ROIs were defined as regions in the lateral and ventral occipital cortex and the inferior IPS region, respectively, revealing higher activations for objects than for noise images (Supplementary Fig. 1b–d)^{20–22}.

VSTM capacity was higher for simple than for complex features (Fig. 2b), with a significant interaction between set size and feature complexity ($n = 7$, $P < 0.001$, analysis of variance, two-tailed; this applies to all subsequent statistical tests except where noted). For simple shape features, *K* values increased linearly with set size ($P < 0.01$), differed among set sizes 2 to 6 ($P < 0.001$) and reached a plateau at set size 4 ($P > 0.18$ for the differences between set sizes 4 and 6). For complex shape features, *K* values did not increase with set size ($P > 0.14$) and reached a plateau at set size 2 ($P > 0.45$ for the differences among set sizes 2 to 6). The maximum VSTM capacity was therefore about four objects for simple shape features and only two objects for complex shape features.

fMRI activations in the superior IPS and LOC tracked the capacity

¹Department of Psychology, Yale University, New Haven, Connecticut 06520-8205, USA.

estimates in behavioural performance, increasing with set size for simple shape features ($P < 0.01$), but not for complex shape features ($P > 0.19$) (Fig. 2b). Although their activations differed among set sizes 2 to 6 for simple shape features ($P < 0.05$), they did not differ for complex shape features ($P > 0.54$), with significant interactions between set size and feature complexity ($P < 0.05$). Moreover, for simple shape features, the increase in activations with set size in these brain regions correlated with the increase in K values ($P < 0.01$). In contrast, regardless of feature complexity, activations in the inferior IPS increased from set sizes 1 to 4 ($P < 0.01$) and reached a plateau at set size 4 ($P > 0.17$ for the differences between set sizes 4 and 6), with no interaction between set size and feature complexity ($P > 0.19$) (Fig. 2b). Thus, whereas activations in the inferior IPS tracked a fixed

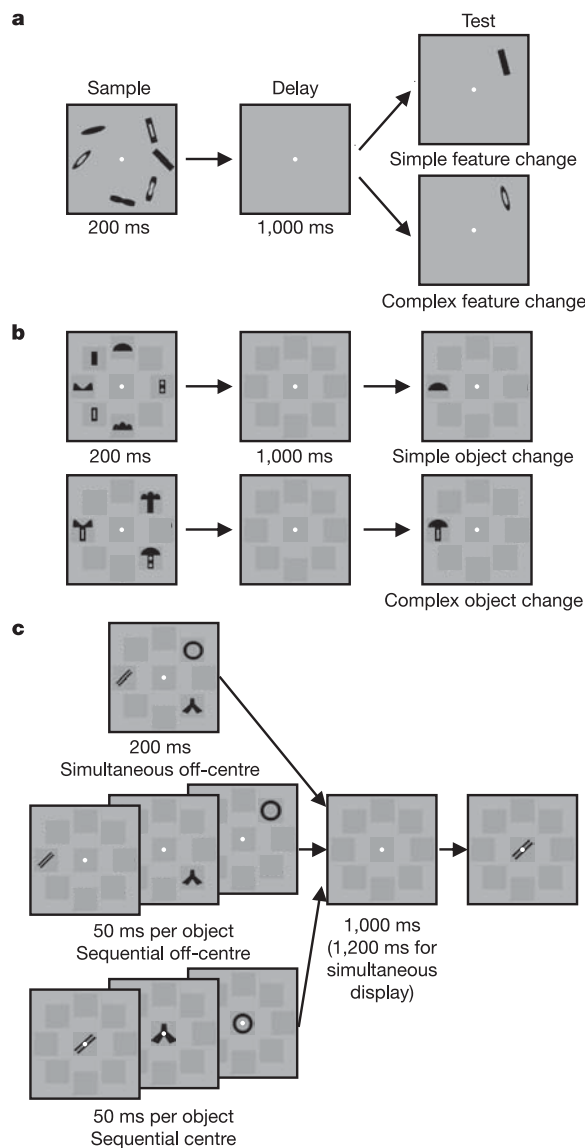


Figure 1 | Example trials from experiments 1, 2 and 4. **a**, In experiment 1, observers viewed the sample displays that consisted of a variable number of black, elongated shapes arranged in a circle, and after a brief blank delay they detected a possible change in the test display at the probed location. In the test display, either a simple feature (presence or absence of the hole in the centre) or a complex shape feature (shape outline) could change. **b**, Experiment 2 was identical to experiment 1, except that each complex object was constructed by attaching two simple objects together. **c**, In experiment 4, observers viewed a variable number of different black objects simultaneously, sequentially at different off-centre locations, or sequentially at the same centre location, and judged whether the probe object in the test display was present in the sample display (sequence).

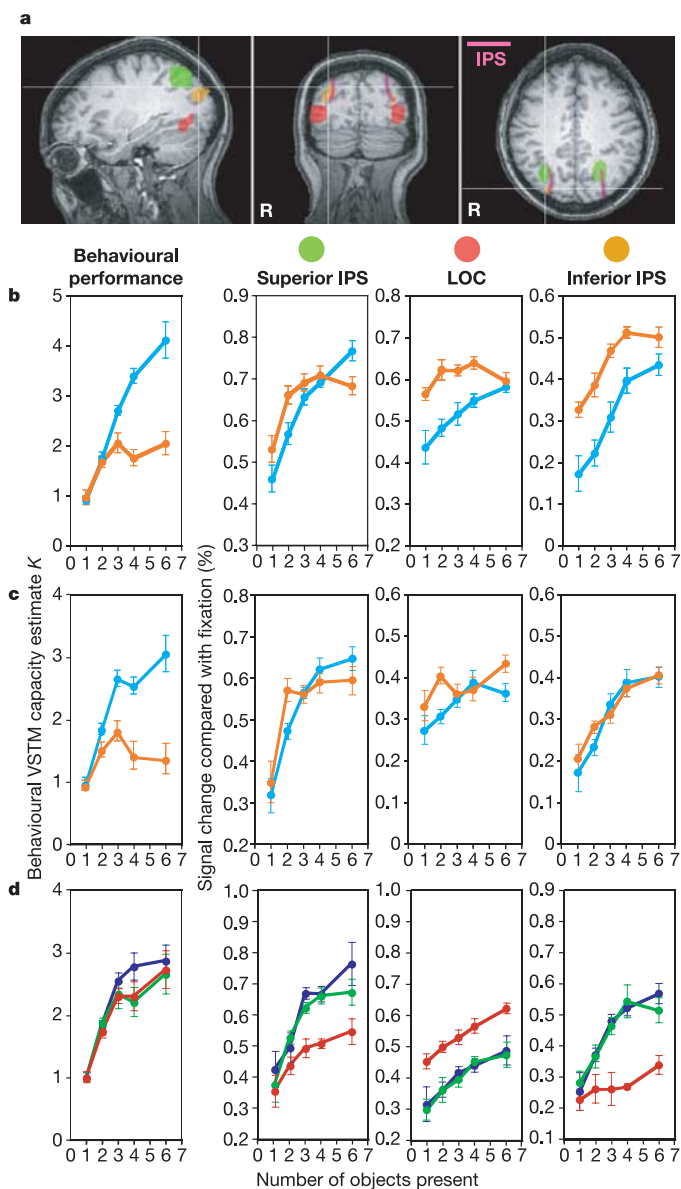


Figure 2 | Results from experiments 1, 2 and 4. **a**, Superior IPS (green), LOC (red) and inferior IPS (orange) ROIs from an example observer in experiment 1. The IPS itself is highlighted in magenta. The averaged Talairach coordinates (x, y, z) for the ROIs are as follows: right/left, $+23/-21 -52/-66 +45/+42$ for the superior IPS, $+26/-21 -80/-85 +30/+26$ for the inferior IPS with off-centre presentation, $+26/-25 -65/-70 +34/+29$ for the inferior IPS with centred presentation (experiment 4), $+42/-44 -67/-69 +3/+8$ for the LOC with off-centre presentation, and $+43/-41 -67/-71 -5/-1$ for the LOC with centred presentation (Experiment 4). **b-d**, Averaged behavioural results and peak responses from the ROIs examined in experiments 1 (**b**), 2 (**c**) and 4 (**d**) (means $\pm$ within-subject s.e.m.). In experiment 1, VSTM capacity was about four objects for simple shape features (cyan) and only two objects for complex shape features (orange). In experiment 2, VSTM capacity was about three for simple objects (cyan) and 1.5 for complex objects (orange). In both experiments, whereas a fixed object capacity was observed in the inferior IPS, a variable object capacity tracking the number of objects held in VSTM was observed in the superior IPS and the LOC as object complexity changed. In experiment 4, VSTM capacity was about three objects in all three display conditions: blue, simultaneous off-centre; green, sequential off-centre; red, sequential centred. Whereas responses in the superior IPS and the LOC tracked VSTM capacity in all three display conditions, those in the inferior IPS did so only when objects appeared in different spatial locations. When objects appeared sequentially at the same central location, however, IPS responses did not vary with set size.

number of objects regardless of object complexity, those in the superior IPS and LOC followed the actual number of objects held in VSTM as object feature complexity changed. Our left superior and inferior IPS ROIs coincide with the region whose activity has been shown to correlate with individual differences in VSTM capacity¹⁸.

Because the encoding of simple shape features required a binary discrimination of hole presence or absence at the centre of each shape, observers could have successfully performed the VSTM task by encoding and remembering only the hole-present (or the hole-absent) shapes without retaining features from the other shapes. This potential grouping strategy could therefore have inflated the estimated VSTM capacity and obscured the corresponding brain activations. Although this strategy would only have reduced the differences between set sizes, inconsistent with the pattern of fMRI results reported above, to test this grouping hypothesis directly we conducted a second fMRI experiment that replicated the design of the first experiment with uniquely different simple and complex objects constructed from the same set of shape features (Fig. 1b).

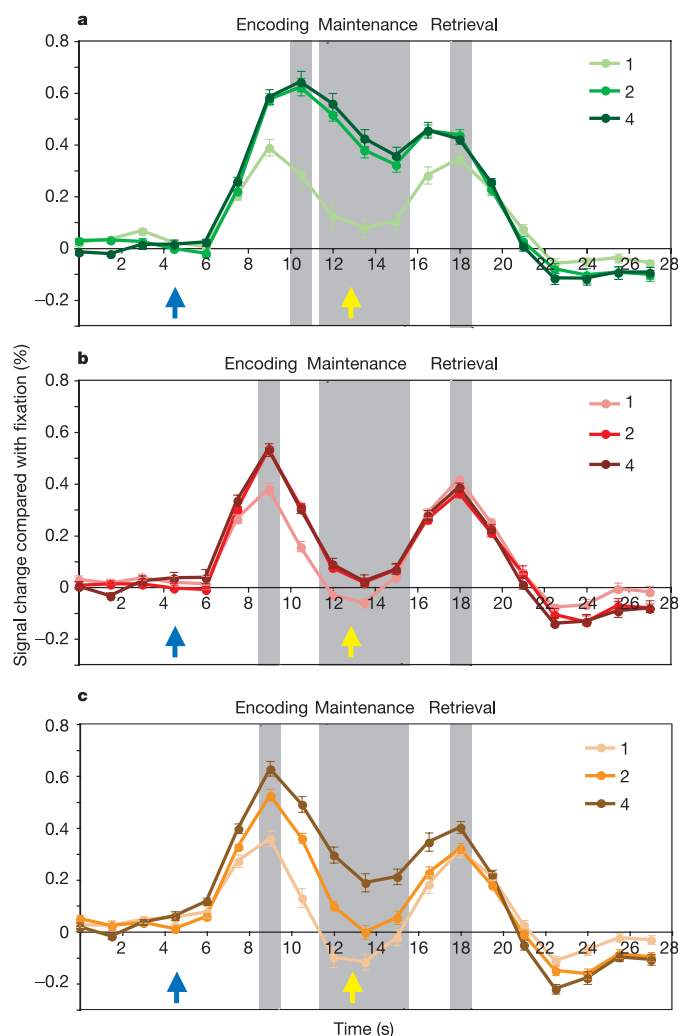


Figure 3 | Results from experiment 3. Activations (means $\pm$ within-subject s.e.m.) from the superior IPS (a), the LOC (b) and the inferior IPS (c). The blue and the yellow arrows indicate onsets of the sample and the test object displays, respectively. Behavioural capacity estimate K values were 0.92 (with a within-subject s.e.m. of 0.14), 1.30 (0.12) and 1.14 (0.22) for set sizes 1, 2 and 4, respectively. Whereas activations in the superior IPS and the LOC followed K values and increased from set sizes 1 to 2 and reached a plateau at set size 2 during both VSTM encoding and maintenance, those in the inferior IPS increased as set size increased from 1 to 4 during both VSTM encoding and maintenance.

As expected, behavioural VSTM capacity was higher for simple than for complex objects (Fig. 2c), with a significant interaction between set size and object complexity ($n = 10$, $P < 0.01$). For simple objects, K values increased with set size ($P < 0.05$), differed among set sizes 2 to 6 ($P < 0.01$), and reached a plateau at set size 3 with K values about 3 objects ($P > 0.29$ for the differences among set sizes 3, 4 and 6). For complex objects, K values did not increase with set size ($P > 0.62$) and reached a plateau at set size 2 with K values about 1.5 objects ($P > 0.49$ for the differences among set sizes 2 to 6). The maximum VSTM capacity was therefore about three objects for simple objects and about 1.5 objects for complex objects.

Replicating the first experiment, superior IPS and LOC activations differed among set sizes 2 to 6 for simple objects ($P < 0.05$), but not for complex objects ($P > 0.73$ for superior IPS, and $P > 0.067$ for LOC; the marginally significant difference was driven by the inexplicably high activation for set size 6 compared with otherwise similar activations among set sizes 2 to 4, $P > 0.28$). The interactions between set size and object complexity in these areas were significant ($P < 0.05$). Moreover, for simple objects, the increase in activations with set size in these brain regions correlated with the increases in K values ($P < 0.05$). In contrast, regardless of object complexity, activations in the inferior IPS increased from set sizes 1 to 4 ($P < 0.001$) and reached a plateau at set size 4 ($P > 0.16$ for the differences between set sizes 4 and 6), with no interaction between set size and object complexity ($P > 0.56$) (Fig. 2c). Thus, even after grouping cues were removed, the results of the second experiment mirrored those of the first experiment, showing a fixed object capacity in the inferior IPS and variable, complexity-dependent object capacity in the superior IPS and the LOC.

It is possible that the lower VSTM capacity for the complex objects was due to perceptual processing limitations rather than memory limitations. To test this hypothesis we conducted a behavioural experiment that provided more time for perceptual processing. Using the complex objects from the second experiment, we found that VSTM capacity did not improve whether the encoding time for four objects was 200 or 500 ms ($n = 6$, $P > 0.82$; $K = 1.90$ and $K = 1.85$, respectively) or whether four objects were presented simultaneously for 200 ms or sequentially two at a time for 200 ms with a 500-ms blank interval in between. In fact, simultaneous presentation was better than sequential presentation ($n = 6$, $P < 0.05$; $K = 1.90$ and $K = 1.60$, respectively). These results indicate that the reduced VSTM capacity for complex objects was due to limitations in VSTM rather than perception.

To dissociate whether brain activations observed so far reflect VSTM encoding, maintenance, or retrieval and comparison, in a third fMRI experiment we presented either one, two or four complex objects from the second experiment (Fig. 1b) and asked observers to maintain the objects over an extended 8,300-ms delay^{9,10,17}. Activations in the LOC and the superior IPS differed between set sizes 1 and 2 and between set sizes 1 and 4 ($n = 8$, $P < 0.05$) but not between set sizes 2 and 4 ($P > 0.18$) during both encoding and maintenance (Fig. 3a, b). These differences were largely maintained in the superior IPS during retrieval ($P < 0.05$ between set sizes 1 and 2, $P < 0.079$ between 1 and 4, and $P > 0.64$ between 2 and 4), but not in the LOC ($P > 0.18$ for the overall differences among set sizes). Activations in the inferior IPS differed between set sizes 1 and 2, between 1 and 4, and between 2 and 4 ($P < 0.05$) during both encoding and maintenance, and these differences were partly retained during retrieval (Fig. 3c, $P < 0.092$ between set sizes 1 and 4, $P < 0.05$ between 2 and 4, and $P > 0.87$ between 1 and 2). These results indicate that brain activations observed in the first two fMRI experiments mainly reflected activations during VSTM encoding and maintenance, consistent with a previous finding on the superior IPS¹⁷.

Finally, to examine whether LOC and IPS activations during VSTM tasks track object identity or simply the locations occupied by objects in the display and in memory, we conducted a fourth fMRI experiment that varied whether objects shared the same or different

spatial locations. One to four, or six, different black objects were presented in one of three ways: simultaneously, sequentially at different off-centre locations, or sequentially at the same centre location. The task was to judge whether the probe object in the test display was present in the sample display. Because the probe object always appeared at the centre of the test display, only VSTM for object identities but not their locations was needed for this task (Fig. 1c). For all three presentation conditions, K varied significantly with set size ($n = 7$, $P < 0.01$), increased as set size increased from 1 to 3 ($P < 0.05$) and reached a plateau at set size 3 ($P > 0.46$ for the differences among set sizes 3, 4 and 6), with no differences between the different conditions ($P > 0.92$). VSTM capacity was therefore about three objects for all three presentation conditions.

To account for differences in eccentricity for centred and off-centre presentations, we separately localized the part of the LOC and the inferior IPS that respond to centred and off-centre presentations (Supplementary Fig. 1d). Regardless of whether objects occupied the same or different spatial locations, fMRI responses in the superior IPS and LOC varied with set sizes ($P < 0.06$ for LOC in the simultaneous condition and $P < 0.05$ for all others) and correlated with the corresponding K values ($P < 0.05$) in all three presentation conditions. However, inferior IPS exhibited a similar response pattern only when objects appeared at different spatial locations regardless of simultaneous or sequential presentation ($P < 0.05$). When objects appeared sequentially at the same central location, however, inferior IPS responses did not vary with set size ($P > 0.45$). Overall LOC responses were greater for centred (fovea) than off-centre presentations ($P < 0.073$), whereas the opposite was true for superior IPS and inferior IPS responses ($P < 0.05$ and $P < 0.093$, respectively). These results indicate that the LOC represents the visual objects held in VSTM and some object location information, as indicated by the object location eccentricity effect in Fig. 2d. In contrast, inferior IPS representations during VSTM are more spatial in nature²⁷, automatically indexing a fixed number of objects by means of their locations, even when the encoding of spatial location is not required. Meanwhile, responses in the superior IPS in this task resembled both those of the LOC (with responses correlating to the number of objects held in VSTM) and those of the inferior IPS (with a lower overall response for centred than off-centre presentations), indicating that both object identity and some location information might be represented in this brain area during VSTM²⁷.

Together, our data from four fMRI experiments indicate dissociable neural mechanisms in the superior and the inferior IPS and the LOC that work in parallel to support VSTM during visual object encoding and maintenance. Regardless of object complexity, inferior IPS representations are limited by a fixed number of objects at different spatial locations, indicating a possible spatial indexing mechanism that maintains spatial attention over a fixed number of objects at different spatial locations during both encoding and maintenance of VSTM. Consistent with its role in attention^{28,29}, this mechanism may select and determine the maximum number of objects that can be maintained in VSTM and possibly other capacity-limited visual tasks such as subitizing and multiple object tracking^{25,30}. In contrast, the representations in the LOC and the superior IPS are not limited by a fixed number of objects, but rather by object complexity and the total amount of visual information encoded, indicating that these brain areas participate in the detailed representation of visual objects in VSTM during both encoding and maintenance. Thus, VSTM capacity is determined both by a fixed number of objects and by object complexity. These findings not only clarify how parietal and occipital cortices mediate VSTM but also help to resolve a central debate over what determines VSTM capacity²⁻⁵.

METHODS

In experiment 1, nine observers viewed elongated and tilted black objects in a sample display and, after a brief delay, detected a shape feature change in the test display at the probed location in 50% of the trials (Fig. 1a). In different task

blocks, observers encoded either simple shape features (the presence or absence of a hole in the shape) or complex shape features (shape outlines). In experiment 2, each complex object was constructed by attaching a cap-like simple object to a stem-like simple object⁴. Twelve observers encoded either simple or complex objects (Fig. 1b). In experiment 3, nine observers retained the complex objects used in experiment 2 for an extended memory delay period. In experiment 4, eight observers encoded objects that were presented simultaneously, sequentially at different off-centre locations, or sequentially at the same centre location, with the probe object always appearing at the same centre location (Fig. 1c).

MRI data were acquired from a Siemens Trio 3T scanner and analysed with BrainVoyager QX 1.3 (<http://www.brainvoyager.com>). After defining the ROIs separately in each observer, we overlaid the ROIs onto the MRI data from our VSTM experiment and extracted time courses from each observer. The resulting peak responses were then averaged across observers. Behavioural data collected during the scans were transformed to VSTM capacity estimates by using Cowan's K formula²⁵: $K = (\text{hit rate} + \text{correct rejection rate} - 1) \times (\text{display set size})$.

More details of the experiments are available in Supplementary Methods.

Received 29 July; accepted 26 September 2005.

Published online 28 December 2005.

- Phillips, W. A. On the distinction between sensory storage and short-term visual memory. *Percept. Psychophys.* **16**, 283–290 (1974).
- Pashler, H. Familiarity and visual change detection. *Percept. Psychophys.* **44**, 369–378 (1988).
- Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
- Xu, Y. Encoding colour and shape from different parts of an object in visual short-term memory. *Percept. Psychophys.* **64**, 1260–1280 (2002).
- Alvarez, G. A. & Cavanagh, P. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol. Sci.* **15**, 106–111 (2004).
- Fuster, J. M. & Jervey, J. P. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *J. Neurosci.* **2**, 361–375 (1982).
- Goldman-Rakic, P. S. In *Handbook of Physiology: The Nervous System, Higher Functions of the Brain* (eds Mountcastle, V. B., Plum, F. & Geiger, S. R.) 373–417 (American Physiological Society, Bethesda, Maryland, 1987).
- Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
- Cohen, J. D. et al. Temporal dynamics of brain activation during a working memory task. *Nature* **386**, 604–608 (1997).
- Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386**, 608–611 (1997).
- Desimone, R. Neural mechanisms for visual memory and their role in attention. *Proc. Natl Acad. Sci. USA* **93**, 13494–13499 (1996).
- Smith, E. E. & Jonides, J. Neuroimaging analyses of human working memory. *Proc. Natl Acad. Sci. USA* **95**, 12061–12068 (1998).
- Pessoa, L., Gutierrez, E., Bandettini, P. A. & Ungerleider, L. G. Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron* **35**, 975–987 (2002).
- Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **9**, 415–423 (2003).
- Jonides, J., Lacey, S. C. & Nee, D. E. Processes of working memory in mind and brain. *Curr. Dir. Psychol. Sci.* **14**, 2–5 (2005).
- Pasternak, T. & Greenlee, M. W. Working memory in primate sensory systems. *Nature Rev. Neurosci.* **6**, 97–107 (2005).
- Todd, J. J. & Marois, R. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* **428**, 751–754 (2004).
- Todd, J. J. & Marois, R. Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cogn. Affect. Behav. Neurosci.* **6**, 144–155 (2005).
- Vogel, E. K. & Machizawa, M. G. Neural activity predicts individual differences in visual working memory capacity. *Nature* **428**, 748–751 (2004).
- Malach, R. et al. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl Acad. Sci. USA* **92**, 8135–8139 (1995).
- Grill-Spector, K., Kushnir, T., Edelman, S., Itzhak, Y. & Malach, R. Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron* **21**, 191–202 (1998).
- Kourtzi, Z. & Kanwisher, N. Cortical regions involved in perceiving object shape. *J. Neurosci.* **20**, 3310–3318 (2000).
- Grill-Spector, K., Kushnir, T., Hendler, T. & Malach, R. The dynamics of object-selective activation correlate with recognition performance in human. *Nature Neurosci.* **3**, 837–843 (2000).
- Awh, E. & Jonides, J. Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* **10**, 433–437 (1999).
- Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001).

26. Talairach, J. & Tournoux, P. *Co-Planar Stereotaxic Atlas of the Human Brain* (transl. Rayport, M) (Thieme Medical, New York, 1988).
27. Sereno, M. I., Pitzalis, S. & Martinez, A. Mapping of contralateral space in retinotopic coordinates by a parietal cortical area in human. *Science* **294**, 1350–1354 (2001).
28. Kanwisher, N. & Wojciulik, E. Visual attention: insights from brain imaging. *Nature Rev. Neurosci.* **1**, 91–100 (2000).
29. Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nature Rev. Neurosci.* **3**, 215–229 (2002).
30. Culham, J. C. *et al.* Attention response functions: Characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron* **32**, 737–745 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Widders for assistance in MRI scanning, and S. J. Luck and R. Marois for comments on earlier versions of this manuscript. This research was supported by an NIH grant to M.M.C. and in part by an NSF grant to Y.X.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.X. (yaoda.xu@yale.edu).

LETTERS

Mitoferrin is essential for erythroid iron assimilation

George C. Shaw^{1†}, John J. Cope^{1†}, Liangtao Li², Kenneth Corson¹, Candace Hersey³, Gabriele E. Ackermann^{1†}, Babette Gwynn⁴, Amy J. Lambert⁴, Rebecca A. Wingert^{3†}, David Traver^{3†}, Nikolaus S. Trede^{3†}, Bruce A. Barut³, Yi Zhou³, Emmanuel Minet¹, Adriana Donovan³, Alison Brownlie^{3†}, Rena Balzan⁵, Mitchell J. Weiss⁶, Luanne L. Peters⁴, Jerry Kaplan², Leonard I. Zon³ & Barry H. Paw¹

Iron has a fundamental role in many metabolic processes, including electron transport, deoxyribonucleotide synthesis, oxygen transport and many essential redox reactions involving haemoproteins and Fe–S cluster proteins. Defective iron homeostasis results in either iron deficiency or iron overload¹. Precise regulation of iron transport in mitochondria is essential for haem biosynthesis², haemoglobin production and Fe–S cluster protein assembly^{3,4} during red cell development. Here we describe a zebrafish mutant, *frascati* (*frs*)⁵, that shows profound hypochromic anaemia and erythroid maturation arrest owing to defects in mitochondrial iron uptake. Through positional cloning, we show that the gene mutated in the *frs* mutant is a member of the vertebrate mitochondrial solute carrier family (SLC25)⁶ that we call *mitoferrin* (*mfrn*). *mfrn* is highly expressed in fetal and adult haematopoietic tissues of zebrafish and mouse. Erythroblasts generated from murine embryonic stem cells null for *Mfrn* (also known as *Slc25a37*) show maturation arrest with severely impaired incorporation of ⁵⁵Fe into haem. Disruption of the yeast *mfrn* orthologues, *MRS3* and *MRS4*, causes defects in iron metabolism and mitochondrial Fe–S cluster biogenesis^{7–10}. Murine *Mfrn* rescues the defects in *frs* zebrafish, and zebrafish *mfrn* complements the yeast mutant, indicating that the function of the gene may be highly conserved. Our data show that *mfrn* functions as the principal mitochondrial iron importer essential for haem biosynthesis in vertebrate erythroblasts.

The zebrafish (*Danio rerio*) is a powerful model system in which to study vertebrate development and heritable diseases in humans, in particular haematological disorders¹¹. The *frs* mutation is one of the genetic complementation groups of zebrafish blood mutants, characterized by embryonic hypochromic anaemia⁵. Five independent *frs* recessive alleles have been identified from the Tübingen 1996 (ref. 5) and 2000 genetic screens: *tg221*, *tg280*, *tm130*, *tq223* and *ij001*. A deficiency allele, *spadetail* (*spt*^{b333})¹², has been identified by genetic non-complementation that contains a large multigenic deletion encompassing the *spt* (*tbx16*)¹² locus and adjacent genes, including the *frs* locus.

Embryonic erythropoiesis is initially normal at 24 h postfertilization (hpf), but increasingly severe anaemia is detectable by 36 hpf (Fig. 1a). The few *frs* mutants that survive to juvenile stage at

2 months postfertilization are pale and growth-retarded and show cardiomegaly, suggestive of high-output heart failure from severe chronic anaemia (Fig. 1a). Forward and side scatter of zebrafish haematopoietic cells by flow cytometry, which distinguishes different cell lineages¹³, revealed a paucity of mature erythrocytes in *frs* zebrafish (Supplementary Fig. 1a). Histological examination of the *frs* haematopoietic cells showed a profound differentiation arrest as immature erythroblasts (Fig. 1b).

We mapped the *frs* locus, using bulk segregant analysis with simple sequence length polymorphism (SSLP) markers, to zebrafish linkage group 8 (Fig. 1c). A closely linked marker, *z-9279*, was determined to be ~0.3 cM away from the *frs* locus (27 recombinants per 8,080 meioses). This marker was used to isolate large genomic bacterial artificial chromosome (BAC) clones and to assemble a physical contig spanning the region. One gene, *fc73c08*, annotated in the critical interval (*Zv5_scaffold 431.1*), was homologous to mitochondrial solute transporters (SLC25)⁶ in *Schizosaccharomyces pombe* (SPAC8C9.12C). Refined mapping showed that the breakpoint of the *spt*^{b333} deletion transects this candidate gene, thereby inactivating its function (Supplementary Fig. 1b).

From adult zebrafish kidney, the site of definitive haematopoiesis in teleosts, we recovered a full-length zebrafish cDNA clone of ~5 kb for *fc73c08*, encoding an open reading frame of 333 amino acids with a predicted polypeptide size of ~37 kDa. Missense mutations in the candidate gene were identified in all of the five *frs* alleles: *tq223* (L46R), *tm130* (M61K), *ij001* (L144H), *tg280* (R158W) and *tg221* (Y207D). These genetic studies strongly argue that the candidate gene, *fc73c08*, is the gene mutated in *frs*.

We refer to the gene mutated in *frs* mutants as *mitoferrin* (*mfrn*; *slc25a37*). Members of this SLC25 solute carrier family form a group of structurally related proteins, which are expressed in the inner mitochondrial membrane of eukaryotes⁶. SLC25 carriers shuttle a diverse range of metabolites and cofactors across the inner mitochondrial membrane. Defects in members of the mitochondrial carrier family have been described for human genetic disorders, including adult-onset type II citrullinaemia¹⁴ (*citrin*; also known as *SLC25A13*) and congenital microcephaly¹⁵ (*deoxynucleotide carrier*; *SLC25A19*).

Whole-mount *in situ* hybridization showed that *mfrn* was highly

¹Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. ²Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA. ³Howard Hughes Medical Institute, Stem Cell Program and Division of Hematology-Oncology, Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴The Jackson Laboratory, Bar Harbor, Maine 04609, USA. ⁵Department of Physiology and Biochemistry, University of Malta, Msida MSD 06, Malta. ⁶Division of Hematology, Department of Pediatrics, Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. [†]Present addresses: Ohio State University College of Medicine, Columbus, Ohio 43210, USA (G.C.S.); University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, New York 14214, USA (J.J.C.); Kinderspital Zürich, Zürich 8032, Switzerland (G.E.A.); Massachusetts General Hospital, Boston, Massachusetts 02114, USA (R.A.W.); University of California, San Diego, California 92093, USA. (D.T.); University of Utah, Salt Lake City, Utah 84112, USA (N.S.T.); Xenon Pharmaceuticals, Burnaby, British Columbia V5G 4W8, Canada (A.B.).

expressed in the intermediate cell mass (ICM), a tissue equivalent to the extra-embryonic yolk-sac blood islands in mammals (Fig. 1d). The specificity of the *in situ* hybridization was confirmed by the absence of detectable transcripts in the *frs* deficiency allele *spt*^{b333}, and in *cloche* mutants, which lack haematopoietic progenitors. All *frs* alleles showed normal levels of *mfrn* mRNA in the ICM at 24 hpf, as expected because their missense mutations would not interfere with posttranscriptional mRNA stability. The transcription factor gene, *gata1*, is essential in regulating erythroid terminal maturation in various organisms¹⁶. The developmental expression pattern of *mfrn* parallels the expression pattern of *gata1* in the lateral plate mesoderm

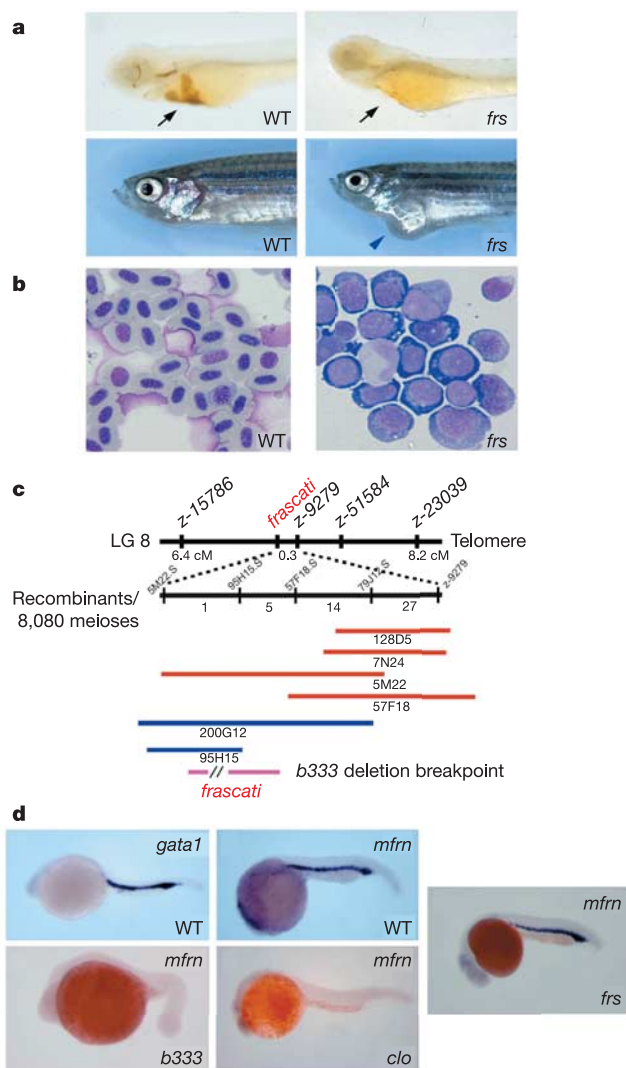


Figure 1 | Analysis of the zebrafish *frs* mutation, the *frs* locus and expression of *mfrn*. **a**, *frs* embryos and juvenile fish are profoundly anaemic. Wild-type and *frs* embryos were stained to detect haemoglobinized cells (top). Juvenile *frs* zebrafish (bottom right) show cardiomegaly (arrowhead) from chronic anaemia. **b**, *frs* blood erythrocytes show maturation arrest (right) at the proerythroblast stage, as compared with mature, nucleated erythrocytes from wild-type fish (left). **c**, Genetic and physical map of the *frs* locus on linkage group 8. The number of recombination events out of 8,080 meioses is noted below the expanded map. Indicated below the critical BAC clones (200G12, 95H15, 5M22) encompassing the *frs* locus is the location of the deletion breakpoint for the deficiency allele *spt*^{b333}. **d**, Zebrafish *mfrn* is expressed in the intra-embryonic blood island ICM. Whole-embryo *in situ* hybridization at 24 hpf shows expression of *gata1* and *mfrn* in the ICM, but an absence of *mfrn* in *spt*^{b333} and *cloche* embryos. *mfrn* RNA expression is normal in *frs* alleles owing to missense mutations.

and ICM (Supplementary Fig. 1c). Expression of *mfrn* is deficient in the zebrafish *gata1* mutant *vlad tepes*¹⁷ (data not shown), suggesting the importance of *gata1* in activating the erythroid differentiation program for *mfrn* expression *in vivo*.

The subcellular localization of the zebrafish Mfrn protein was studied by transient expression of green fluorescence protein (GFP)-tagged complementary DNA constructs in human HEK293 cells. Confocal fluorescence images showed colocalization of the Mfrn transporter with porin, a resident mitochondrial protein, in the same organelle compartment (Fig. 2a).

To verify that *mfrn* is the gene mutated in *frs*, we introduced *mfrn* sense RNA by microinjection into *frs* embryos at the one-cell stage and assayed for restoration of haemoglobinization (Fig. 2b, c). The injected embryos were allowed to develop to 72 hpf, stained with *o*-dianisidine to detect haemoglobinized cells, and genotyped. Half of the genotyped *frs* mutant embryos were robustly rescued (50% rescue efficiency: 25 rescued/50 mutants). Microinjection with a mutant *mfrn* cRNA based on the *frs*^{fr223} allele did not rescue any *frs* embryos (0 rescued/154 injected), indicating the specificity of the rescue assay.

To confirm the loss-of-function phenotype for *mfrn*, an antisense morpholino¹⁸ designed to the splice-donor site of exon 1 was used to induce selective *mfrn* RNA mis-splicing and degradation. Morphant embryos were profoundly anaemic, lacking any detectable *o*-dianisidine-positive cells, thereby mimicking the *frs* phenotype (Fig. 2d). The anaemic phenotype is attributable to loss of *mfrn* function because no detectable *mfrn* transcript was observed in morphant embryos by *in situ* hybridization. The expression of control embryonic β 2-globin was not perturbed by *mfrn* morpholino injection, demonstrating the selectivity of the phenotype. The injection of a control morpholino had no effect on *o*-dianisidine staining or on the level of *mfrn* or β 2-globin mRNA. Together, these data prove that the *frs* mutation is caused by defects in the *mfrn* gene.

A related paralogous gene sharing 65% amino acid identity, which we call *mitoferrin2* (*mfrn2*; *slc25a28*), is present in zebrafish and mammals (Fig. 3a). The zebrafish *mfrn2* locus was mapped to a different linkage group 13 (data not shown), confirming that *mfrn2* is a related but independent gene. Zebrafish Mfrn and Mfrn2 proteins share about 38% amino acid identity to two ancestral transporters in *Saccharomyces cerevisiae*, MRS3 and MRS4 (hereafter referred to MRS3/4), which have been implicated in mitochondrial metal transport^{7–10}. Both human and mouse homologues of yeast MRS3/4 mitochondrial transporters have been reported^{19–21}; however, the biological function of these mammalian proteins is unclear.

We evaluated the tissue and developmental expression patterns of mouse *Mfrn* by northern blot and *in situ* hybridization analyses. *Mfrn* is abundantly expressed in haematopoietic organs, fetal liver, bone marrow and spleen (Fig. 3b, c). In the developing mouse embryo, expression of *Mfrn* is first detected at embryonic day 7.5 (E7.5) in the extra-embryonic yolk sac, coincident with the appearance of blood islands (Fig. 3c). This is analogous to the expression kinetics of zebrafish *mfrn* in the ICM. Later, restricted expression of *Mfrn* is seen in E14.5 fetal liver, the primary source of erythrocyte production in mid-gestation (Fig. 3c). We found that *Mfrn* mRNA is strongly upregulated during chemically induced maturation of Friend mouse erythroleukaemia (MEL) and G1ER cells²² (Supplementary Fig. 2a, b). Given its abundant expression pattern in haematopoietic tissues, mouse *Mfrn* on chromosome 14 is the functional orthologue of the zebrafish *mfrn* gene. Microinjection of mouse *Mfrn* cRNA into *frs* embryos rescued the anaemia phenotype with the same efficiency as the zebrafish clone (58% rescue efficiency: 53 rescued/91 mutants).

Mouse *Mfrn2* showed ubiquitous expression at low levels in all tissues by northern analysis (Fig. 3b) and by *in situ* hybridization of a mouse embryo at E14.5 (data not shown). The level of *Mfrn2* mRNA did not vary during erythroid maturation of chemically induced MEL cells (Supplementary Fig. 2a). Zebrafish *mfrn2* mRNA was present at low levels throughout the whole embryo at 24 hpf

(Supplementary Fig. 2c). Moreover, neither mouse nor zebrafish *mfrn2* cRNA rescued the anaemia in *frs* zebrafish embryos (zebrafish: 0 rescued/71 injected embryos; mouse: 0 rescued/46 injected embryos). These results indicate a unique, non-redundant function for *mfrn* in vertebrate erythropoiesis.

To develop a cell-based assay to study the biochemical function of *Mfrn* in erythropoiesis, we created an immortalized haematopoietic cell line from mouse embryonic stem (ES) cells as described²³. Characterization of the *Mfrn*^{-/-} haematopoietic cells by polymerase chain reaction with reverse transcription (RT-PCR) showed a complete loss of *Mfrn* mRNA (Supplementary Fig. 3a). After 3 d of growth with recombinant human erythropoietin and interleukin-3, the percentages of *o*-dianisidine-positive cells were as follows (100 cells counted for each): wild-type cells, 15%; heterozygous cells, 7%; and null cells, 0%. These results confirm the inability of the *Mfrn*^{-/-} haematopoietic cells to undergo terminal erythroid maturation.

A defect in iron delivery to the mitochondria could account for the observed phenotype in *frs* zebrafish and in *Mfrn*-null mouse erythroblasts. To assay the incorporation of ⁵⁵Fe into mitochondrial haem²⁴, we metabolically labelled ~2 × 10⁶ wild-type cells with ⁵⁵Fe-saturated transferrin and the haem precursor aminolevulinic acid (ALA), for 6 h. As a control, an equal number of wild-type haematopoietic cells were co-cultivated with succinylacetone, which inhibits the haem synthetic enzyme ALA dehydratase². Wild-type cells efficiently incorporated ⁵⁵Fe into haem; however, cells treated with succinylacetone showed a 90% decrement in their ability to incorporate ⁵⁵Fe into haem (Fig. 4a). *Mfrn*^{-/-} haematopoietic cells were deficient in their ability to incorporate ⁵⁵Fe into haem, indicating that *Mfrn* is required for haemoglobinization.

The yeast *mrs3/4* double mutant (Δ *mrs3/4*) has a biochemical defect in haemoprotein and Fe-S cluster protein biogenesis, which manifests as a growth defect under iron-depleted conditions^{7–10}. The introduction of zebrafish *mfrn* or *mfrn2* efficiently complemented the growth defect of Δ *mrs3/4* mutants in iron-depleted conditions (Fig. 4b). A similar result has been obtained for human *MFRN2* by others¹⁹. Four of the stronger *frs* mutant alleles (*tg221*, *tq223*, *tg280* and *ij001*) did not complement the Δ *mrs3/4* mutation, whereas a hypomorphic allele (*tm130*) complemented the growth defect. The tremendous demand for mitochondrial iron in developing erythroblasts² explains the non-redundant properties of *mfrn* in rescuing anaemic *frs* embryos. By contrast, the basal requirement for mito-

chondrial iron in yeasts and non-erythroid cells is sufficiently provided by *mfrn2*. The deficiency of *MRS3/4* genes decreased mitochondrial aconitase activity by ~50%, whereas introduction of the zebrafish *mfrn* cDNA into the Δ *mrs3/4* host completely restored mitochondrial aconitase activity to normal levels (Supplementary Fig. 3b).

We studied the function of zebrafish *mfrn* as an importer of mitochondrial iron using an assay independent of haem and Fe-S biosynthesis. We stably expressed an unique iron-cofactor-dependent superoxide dismutase (Fe-SOD) from *Escherichia coli*²⁵ in yeast as a surrogate marker for mitochondrial iron homeostasis. Cell lysates from wild-type and Δ *mrs3/4* yeasts were assayed for Fe-SOD activity using a fluorogenic substrate²⁵. The introduction of zebrafish *mfrn* cDNA into Δ *mrs3/4* cells fully restored the activity of *E. coli* Fe-SOD to normal levels as compared with cells transformed with a control vector, which had half the normal activity of Fe-SOD (Supplementary Fig. 4a). Immunoblot analyses showed the presence of Fe-SOD and zebrafish *Mfrn* in the mitochondrial fraction with the resident organelle marker, porin (Supplementary Fig. 4b, c). Together, these results in yeast indicate that zebrafish *Mfrn* has a role in mitochondrial iron importation that is highly conserved.

Our results enable us to propose a model of *Mfrn* function in delivering iron into the mitochondria of developing erythroid cells (Fig. 4c). We postulate that its paralogue, *Mfrn2*, is required for haem synthesis of haemoproteins and Fe-S cluster assembly in non-erythroid cells. The iron delivered by *Mfrn* into the mitochondria, presumably as Fe²⁺ as shown for yeast *MRS3/4* (ref. 8), would then be delivered to ferrochelatase to catalyse Fe²⁺ incorporation into protoporphyrin IX to make haem. Our proposed function for the *mfrn* genes in mitochondrial iron acquisition is consistent with studies of the related *MRS3/4* genes in yeast^{7–10}.

A precedence for two paralogous genes, one expressed exclusively in erythroid tissue and another expressed ubiquitously, as in the *mfrn* genes, can be found in the ALA synthase (*ALAS*) genes in mammals and zebrafish. Analysis of the mouse and zebrafish *mfrn* and *mfrn2* 5'- and 3'-untranslated regions did not identify iron-response elements, which posttranscriptionally regulate various genes involved in iron metabolism¹. The differences in tissue expression for *mfrn* and *mfrn2* could be easily accounted for by the presence of erythroid-specific regulatory *cis*-elements for GATA1, EKLF and p45NF-E2 transcription factors in the *mfrn* promoter, as are present

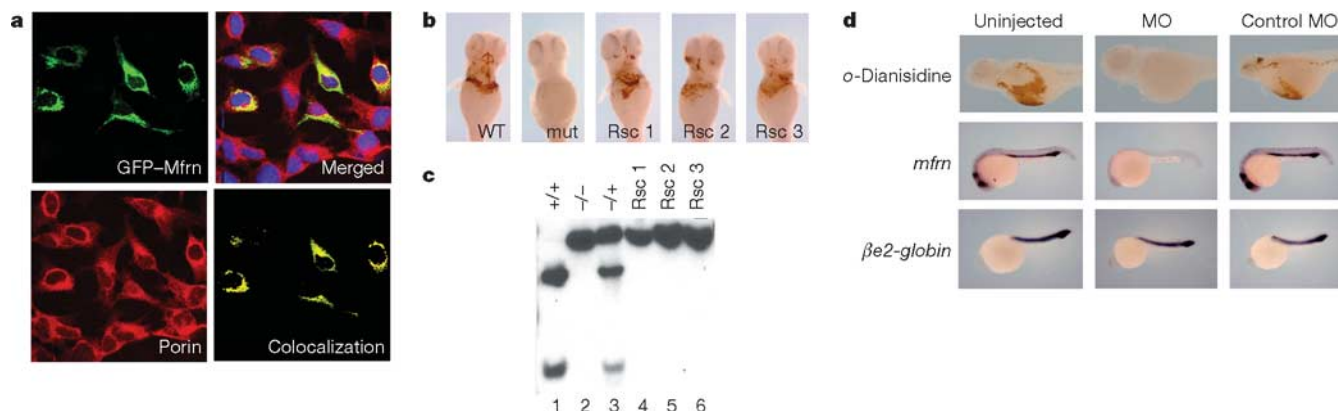


Figure 2 | Mitochondrial localization of *Mfrn* and functional studies of the *mfrn* gene. **a**, Immunolocalization of GFP-tagged *Mfrn* protein to the mitochondria in HEK293 cells. Fluorescence confocal images were obtained from immunostained mitochondrial porin (rhodamine) and GFP-*Mfrn* (fluorescein). Colocalized expression of *Mfrn* in the mitochondria is indicated by the yellow signal (right); nuclei are in blue (DAPI). **b**, Ectopic expression of *mfrn* cRNA rescues the anaemia in *frs* embryos. Control wild-type, *frs* mutant (mut) and rescued *frs* embryos (Rsc 1–3) were stained with *o*-dianisidine to detect haemoglobinized cells. **c**, Control wild-type (lane 1),

mutant (lane 2), heterozygote (lane 3) and rescued *frs* (lanes 4–6) embryos were genotyped by restriction enzyme digestion with *Bsr*I for the *frs*^{*tq223*} mutation. **d**, Morpholino (MO) knockdown of *mfrn* phenocopies the anaemia in wild-type embryos. Uninjected embryos (left) were stained for haemoglobin at 48 hpf or hybridized with *mfrn* or *globin* probes at 24 hpf. Knockdown of *mfrn* (middle) resulted in anaemia and absence of *mfrn* transcripts, without any effect on β e2-*globin* mRNA. Injections with a control morpholino had no effect on erythropoiesis or expression of the two erythroid-specific transcripts (right).

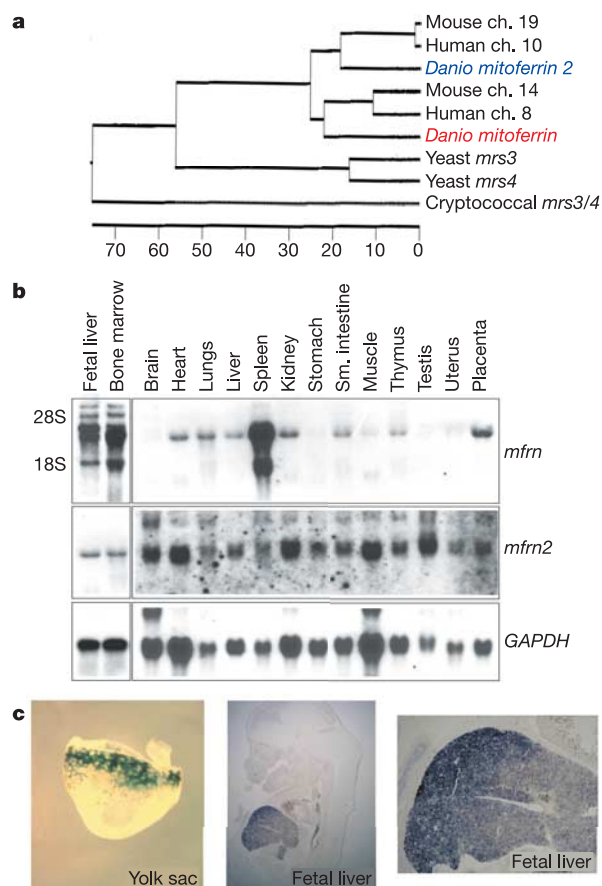


Figure 3 | Phylogenetic relationship and expression pattern of *mfrn* and its paralogue, *mfrn2*. **a**, Phylogenetic dendrogram showing amino acid homology between the various *mfrn* genes. The *D. rerio mfrn* gene (red) is shown clustering with its functional mammalian orthologues. A paralogous gene, *mfrn2* (blue), is shown clustering with its mammalian orthologues on different chromosomes. **b**, Mouse *Mfrn* is highly expressed in haematopoietic organs, fetal liver, adult bone marrow and spleen. *Mfrn2* is ubiquitously expressed at low levels. Equal RNA loading was normalized by expression of *GAPDH*. **c**, *In situ* hybridization detected expression of *Mfrn* in the yolk sac at E7.5 and fetal liver at E14.5. A higher magnification of fetal liver is shown on the far right.

in the *ALAS2* promoter²⁶. Our results for zebrafish *mfrn* expression in *vlad tepes* and murine *Mfrn* expression in G1ER cells are consistent with the dependence of this gene on GATA1 (Supplementary Figs 1c and 2b).

Our results, in conjunction with data from yeast, clearly show that the *mfrn* carrier family is responsible for mitochondrial iron accumulation. Given the marked phenotype in zebrafish and mouse, we predict that *MFRN* on human chromosome 8p21 may be important in human congenital types of anaemia. The discovery of *mfrn* as an essential mitochondrial iron importer provides a powerful genetic tool for exploring human disorders of iron metabolism and erythropoiesis.

METHODS

Zebrafish strains. We recovered the *frs* alleles *tg280*, *tg221*, *tm130*, *tq223* and *ij001* from screens in Tübingen, Germany⁵. The deficiency allele *spt*^{b333} was obtained from M. Halpern¹². We maintained *frs* fish on a standard genetic AB or Tü background. The polymorphic strain used for bulk segregant mapping was SJD obtained from S. Johnson.

Genetic mapping, chromosomal walk and cDNA selection. We used embryo DNA from AB/SJD or Tü/SJD hybrids for linkage analysis to linkage group 8. We genotyped with the SSLP markers *z-15786*, *z-9279*, *z-51584* and *z-23039*. We used marker *z-9279* to isolate clones from the Danikey and CHORI BAC libraries (Sanger Centre, UK) for assembling a physical contig (Zv5_scaffold 431.1-contig

13416.1) through the *frs* locus. One annotated gene, EST *fc73c08*, was used as a probe to screen an adult kidney cDNA library in λZap Express for a full-length cDNA clone, which we refer to as *mitoferlin* (*mfrn*). A zebrafish paralogue, *mitoferlin2* (*mfrn2*), was identified from the GenBank database and obtained from Open Biosystems. The full-length mouse *Mfrn* cDNA was isolated from mouse bone marrow by RT-PCR with primers derived from published sequences²¹. Mouse *Mfrn2* cDNA was obtained from Open Biosystems.

Mutational analysis. We isolated total RNA from wild-type and *frs* embryos at 3–4 days postfertilization (dph) and analysed the cDNA as described²⁷. Primer sequences for amplification and genotyping information by restriction digestion and allele-specific oligonucleotide hybridization are available from the authors on request.

In situ hybridization and o-dianisidine staining. We carried out zebrafish whole-embryo *in situ* hybridization with *mfrn*, *mfrn2* and *gata1* digoxigenin-labelled probes as described⁵. We stained live embryos with o-dianisidine at 2–4 dph as described²⁸.

cRNA injections for expression in zebrafish. We subcloned cDNA clones for zebrafish *mfrn* and *mfrn2*, and mouse *Mfrn* and *Mfrn2* into a pCS2+ vector. We prepared 5'-capped cRNA using the SP6 mMessage mMachine kit (Ambion). Fertilized eggs from *frs*^{tq223} heterozygous pairs were injected at the 1–2-cell stage with roughly 100 pg of cRNA and phenol red dye as a tracer.

Morpholino knockdown in zebrafish. An antisense morpholino¹⁸ (GeneTools) complementary to the splice-donor site of exon 1 was synthesized: 5'-TAAGTTGCATTACCTTGACTGAATC-3'. A morpholino provided by GeneTools was used as a negative control: 5'-CCTCTTACCTCAGTTACAATTATA-3'. Injected (morpholino doses at 0.5–1 mM) and control embryos were either fixed at 24 hpf for whole-embryo *in situ* hybridization or stained with o-dianisidine at 2 dpf (ref. 28).

Antibodies. We generated polyclonal antisera directed against the carboxy-terminal 14 amino acids from zebrafish *mfrn* (C-HESHVQEVSHKTSPT; Genemed Synthesis). Antisera against mitochondrial porin (sc-11433) were purchased from Santa Cruz Biotechnology.

Transient transfection in HEK293 cells and immunohistochemistry. Zebrafish *mfrn* was fused at the amino terminus to GFP in a pEGFP-C3 vector (Clontech). Immunohistochemistry was done essentially as described²⁷ with primary antibody against porin (1:100 dilution) and Texas-red-conjugated secondary antibody (1:200 dilution). Confocal fluorescence images were captured at the Harvard Center for Neurodegeneration.

Flow cytometry. Two-month-old age-matched wild-type and *frs* mutant fish were killed and single-cell suspensions were subjected to flow cytometry using forward and side scatter as described¹³.

Mouse tissue northern analysis. Tissue northern blot was obtained from Seegene. The nylon filters were serially probed with ³²P-labelled mouse *Mfrn*, *Mfrn2* and *GAPDH* probes using standard procedures.

Mouse embryo section in situ hybridization and β-galactosidase staining. C57/BL6 mouse embryos at E14.5 were fixed in paraformaldehyde and sectioned frozen at a thickness of 16 μm. Digoxigenin-labelled *Mfrn* and *Mfrn2* RNA were synthesized and hybridized on sectioned mouse embryos as described²⁹.

Creation of *Mfrn*-deficient haematopoietic cell lines. A gene-trapped ES clone (XB454) was obtained from the UCSF BayGenomics repository. *Mfrn*-null ES clones were derived from the founder ES clone by selection in G418 antibiotic (1.0–2.5 mg ml⁻¹) for 7–10 d as described²³, and their genotype was confirmed by Southern analysis. *Mfrn*-null ES cells were differentiated *in vitro* to form embryoid bodies and subsequently immortalized by co-cultivation with an irradiated NIH3T3 hph-HOX11 producer line as described²³. cDNA was synthesized with Superscript (Invitrogen) and random hexamers. RT-PCR was done with [α-³²P]dCTP (3,000 Ci mmol⁻¹) and primers for mouse *Mfrn* and *GAPDH*.

⁵⁵Fe-metabolic labelling and quantification of haem biosynthesis. ⁵⁵Fe-saturated transferrin was prepared as described³⁰, using ⁵⁵FeCl₃ (76 mCi mg⁻¹, Perkin Elmer). Metabolic labelling of *Mfrn*-deficient and wild-type immortalized haematopoietic cells (~2 × 10⁶) was done as described²⁴. ⁵⁵Fe complexed as haem was quantified in a liquid scintillation counter.

Yeast complementation analysis and aconitase activity assay. Zebrafish cDNAs were subcloned into the pRS426-ADH vector and transformed into the yeast *Δmrs3/4* mutant strain as described⁹. Site-directed mutagenesis of the zebrafish *mfrn* cDNA to recreate the mutations identified in the *frs* alleles was done with a QuikChange kit (Stratagene). Wild-type and *Δmrs3/4* yeast strains transformed with the zebrafish *mfrn* were assayed for mitochondrial aconitase activity as described⁹.

Assay for Fe-SOD activity. Fe-SOD from *E. coli*²⁵ was stably expressed in wild type and *Δmrs3/4* yeast strains. Cell lysates were fractionated on a non-denaturing polyacrylamide gel, and *in situ* Fe-SOD activity staining was done as described²⁵.

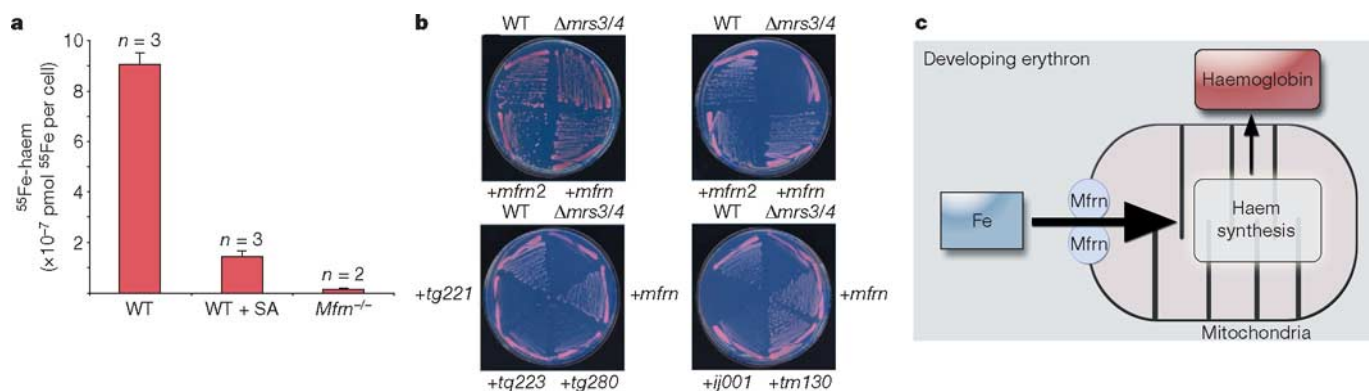


Figure 4 | The function of Mfrn in mitochondrial iron assimilation is conserved. **a**, *Mfrn*-null murine haematopoietic cells are defective in the incorporation of mitochondrial iron for haem biosynthesis. Wild-type, wild-type treated with the porphyrin-synthesis inhibitor succinylacetone (WT + SA), and *Mfrn*^{-/-} haematopoietic cells were incubated with ^{55}Fe -saturated transferrin. Incorporation of ^{55}Fe into haem was quantified by liquid scintillation counting of the cell lysate. Standard errors were calculated from the number of samples indicated above the error bars. **b**, Zebrafish *mfrn* and its paralogue complement the growth defect of the yeast $\Delta\text{mrs3/4}$ mutant in low iron media. Wild-type and $\Delta\text{mrs3/4}$ cells transformed with a control vector or with zebrafish *mfrn* (+*mfrn*) or *mfrn2*

(+*mfrn2*) grew well on iron-repleted media (top left). $\Delta\text{mrs3/4}$ yeast grew poorly on iron-depleted media, but the yeast strain was complemented with either zebrafish *mfrn* gene (top right). The yeast mutant was not complemented by the zebrafish null mutant alleles (+*tg221*, +*tg223*, +*tg280* or +*ij001*), except a hypomorphic (+*tm130*) allele (bottom). **c**, Model of *Mfrn* function in importing iron into the mitochondrial matrix for haem synthesis and haemoglobin production in developing erythroblasts. The iron delivered to the mitochondria by Mfrn would be released to ferrochelatase, catalysing the insertion of Fe^{2+} into protoporphyrin IX to make haem, which is exported to the cytoplasm.

Received 12 August; accepted 5 December 2005.

- Hentze, M. W., Muckenthaler, M. U. & Andrews, N. C. Balancing acts: molecular control of mammalian iron metabolism. *Cell* **117**, 285–297 (2004).
- Napier, I., Ponka, P. & Richardson, D. R. Iron trafficking in the mitochondria: novel pathways revealed by disease. *Blood* **105**, 1867–1874 (2005).
- Lill, R. & Mühlenhoff, U. Iron-sulfur-protein biogenesis in eukaryotes. *Trends Biochem. Sci.* **30**, 133–141 (2005).
- Rouault, T. A. & Tong, W. H. Iron-sulphur cluster biogenesis and mitochondrial iron homeostasis. *Nature Rev. Mol. Cell Biol.* **6**, 345–351 (2005).
- Ransom, D. G. et al. Characterization of zebrafish mutants with defects in embryonic hematopoiesis. *Development* **123**, 311–319 (1996).
- Wohlrab, H. The human mitochondrial transport protein family: identification and protein regions significant for transport function and substrate specificity. *Biochim. Biophys. Acta* **1709**, 157–168 (2005).
- Foury, F. & Roganti, T. Deletion of the mitochondrial carrier genes *MRS3* and *MRS4* suppresses mitochondrial iron accumulation in a yeast frataxin-deficient strain. *J. Biol. Chem.* **277**, 24475–24483 (2002).
- Mühlenhoff, U. et al. A specific role of the yeast mitochondrial carriers *Mrs3/4p* in mitochondrial iron acquisition under iron-limiting conditions. *J. Biol. Chem.* **278**, 40612–40620 (2003).
- Li, L. & Kaplan, J. A mitochondrial-vacuolar signalling pathway in yeast that affects iron and copper metabolism. *J. Biol. Chem.* **279**, 33653–33661 (2004).
- Zhang, Y., Lyver, E. R., Knight, S. A. B., Lesuisse, E. & Dancis, A. Frataxin and mitochondrial carrier proteins, *Mrs3p* and *Mrs4p*, cooperate in providing iron for heme synthesis. *J. Biol. Chem.* **280**, 19794–19807 (2005).
- Shafizadeh, E. & Paw, B. H. Zebrafish as a model of human hematologic disorders. *Curr. Opin. Hematol.* **11**, 255–261 (2004).
- Griffin, K. J., Amacher, S. L., Kimmel, C. B. & Kimmel, D. Molecular identification of *spadetail* regulation of zebrafish trunk and tail mesoderm formation by the T-box genes. *Development* **125**, 3379–3388 (1998).
- Traver, D. et al. Transplantation and *in vivo* imaging of multilineage engraftment in zebrafish bloodless mutants. *Nature Immunol.* **4**, 1238–1246 (2003).
- Kobayashi, K. et al. The gene mutated in adult-onset type II citrullinaemia encodes a putative mitochondrial carrier protein. *Nature Genet.* **22**, 159–163 (1999).
- Rosenberg, M. J. et al. Mutant deoxynucleotide carrier is associated with congenital microcephaly. *Nature Genet.* **32**, 175–179 (2002).
- Cantor, A. B. & Orkin, S. H. Hematopoietic development: a balancing act. *Curr. Opin. Genet. Dev.* **11**, 513–519 (2001).
- Lyons, S. E. et al. A nonsense mutation in zebrafish *gata1* causes the bloodless phenotype in *vlad tepes*. *Proc. Natl Acad. Sci. USA* **99**, 5454–5459 (2002).
- Nasevicius, A. & Ekker, S. C. Effective targeted gene 'knockdown' in zebrafish. *Nature Genet.* **26**, 216–220 (2000).
- Li, F. Y. et al. Characterization of a novel human putative mitochondrial transporter homologous to the yeast mitochondrial RNA splicing proteins 3 and 4. *FEBS Lett.* **494**, 79–84 (2001).
- Li, F. Y., Leibiger, B., Leibiger, I. & Larsson, C. Characterization of a putative murine mitochondrial transporter homology to hMRS3/4. *Mamm. Genome* **13**, 20–23 (2002).
- Li, Q. Z. et al. Rapid decrease of RNA level of a novel mouse mitochondrial solute carrier protein (Mscp) gene at 4–5 weeks of age. *Mamm. Genome* **12**, 830–836 (2001).

- Gregory, T. et al. GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-x_L expression. *Blood* **94**, 87–96 (1999).
- Cantor, A. B., Katz, S. G. & Orkin, S. H. Distinct domains of the GATA-1 cofactor FOG-1 differentially influence erythroid versus megakaryocytic maturation. *Mol. Cell. Biol.* **22**, 4268–4279 (2002).
- Sturrock, A., Alexander, J., Lamb, J., Craven, C. M. & Kaplan, J. Characterization of a transferrin-independent uptake system for iron in HeLa cells. *J. Biol. Chem.* **265**, 3139–3145 (1990).
- Balzan, R., Bannister, W. H., Hunter, G. J. & Bannister, J. V. *Escherichia coli* iron superoxide dismutase targeted to the mitochondria of yeast cells protects the cells against oxidative stress. *Proc. Natl Acad. Sci. USA* **92**, 4219–4223 (1995).
- Surinya, K. H., Cox, T. C. & May, B. K. Transcriptional regulation of the human erythroid 5-aminolevulinic synthase gene. *J. Biol. Chem.* **272**, 26585–26594 (1997).
- Paw, B. H. et al. Cell-specific mitotic defect and dyserythropoiesis associated with erythroid band 3 deficiency. *Nature Genet.* **34**, 59–64 (2003).
- Iuchi, I. & Yamamoto, M. Erythropoiesis in the developing rainbow trout, *Salmo gairdneri* irideus: histological and immunochemical detection of erythropoietic organs. *J. Exp. Zool.* **226**, 409–417 (1983).
- McKee, A. et al. A genome-wide *in situ* hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev. Biol.* **5**, 14 (2005).
- Roy, C. N., Penny, D. M., Feder, J. N. & Enns, C. A. The hereditary hemochromatosis protein, HFE, specifically regulates transferrin-mediated iron uptake in HeLa cells. *J. Biol. Chem.* **274**, 9022–9028 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Lawrence, A. Walker and C. Belair for zebrafish animal husbandry; M. Halpern for *spt^{b333}* zebrafish; S. Johnson for SJD zebrafish; W. Shu Wu for mouse bone marrow cDNA; C. Roy for advice on preparing ^{55}Fe -saturated transferrin; M. Kaku and Y. Fujiwara for advice on ES cell culture and selection; H. Wohlrab for advice on SLC25 biochemistry; K. Dooley for the *frs^{ij001}* allele from the Tübingen 2000 screen; M. Ocaña for help with confocal fluorescence images; S. Dallaire and C. Lee for karyotyping the *Mfrn*-null ES cells; A. B. Cantor for advice on generating Hox-11-immortalized haematopoietic cells; and E. Shafizadeh, S.-K. Choe, C. Burns, Y. Houvras, D. Langenau, W. Tse, H. Wolhrab, J. Kanki and G. M. Shaw for critically reading the manuscript. This work was supported in part by the William Randolph Hearst Foundation (B.H.P.); the March of Dimes Birth Defects Foundation Basil O'Connor Award (B.H.P.); the Swiss National Science Foundation (G.E.A.); the Belgian National Research Fund (E.M.); the NIH (L.L.P., L.I.Z., J.K., M.J.W., B.H.P.); and the Howard Hughes Medical Institute (L.I.Z.).

Author Information Sequences have been deposited in GenBank as follows: Zebrafish *mfrn* (*slc25a37*; DQ112224, DQ112225), Zebrafish *mfrn2* (*slc25a28*; BC054641), mouse *Mfrn* (*Slc25a37*; AF361699), mouse *Mfrn2* (*Slc25a28*; BC025908.1), human *MFRN* (MSCP; *SLC25A37*; NM_016612) and human *MFRN2* (*SLC25A28*; BC058937). Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.H.P. (bpaw@rics.bwh.harvard.edu).

How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP

Agnidipta Ghosh^{1†}, Gerrit J. K. Praefcke^{1†}, Louis Renault², Alfred Wittinghofer¹ & Christian Herrmann^{1†}

Interferons are immunomodulatory cytokines that mediate anti-pathogenic and anti-proliferative effects in cells¹. Interferon- γ -inducible human guanylate binding protein 1 (hGBP1) belongs to the family of dynamin-related large GTP-binding proteins², which share biochemical properties not found in other families of GTP-binding proteins such as nucleotide-dependent oligomerization and fast cooperative GTPase activity³. hGBP1 has an additional property by which it hydrolyses GTP to GMP in two consecutive cleavage reactions^{4,5}. Here we show that the isolated amino-terminal G domain of hGBP1 retains the main enzymatic properties of the full-length protein and can cleave GDP directly. Crystal structures of the N-terminal G domain trapped at successive steps along the reaction pathway and biochemical data reveal the molecular basis for nucleotide-dependent homodimerization and cleavage of GTP. Similar to effector binding in other GTP-binding proteins, homodimerization is regulated by structural changes in the switch regions. Homodimerization generates a conformation in which an arginine finger and a serine are oriented for efficient catalysis. Positioning of the substrate for the second hydrolysis step is achieved by a change in nucleotide conformation at the ribose that keeps the guanine base interactions intact and positions the β -phosphates in the γ -phosphate-binding site.

Guanylate-binding proteins constitute a highly conserved vertebrate-specific interferon- γ -induced family^{2,6}. hGBP1 provides resistance against proliferation of vesicular stomatitis virus and encephalomyocarditis virus⁷, and antagonizes the proliferation and angiogenic activity of endothelial cells on inflammation^{8,9}. The slow intrinsic GTPase activity of hGBP1 is strongly stimulated (up to 100 min⁻¹) by GTP-dependent dimerization³. A unique feature of GBPs is their ability to bind GMP, GDP and GTP with similar affinity and to hydrolyse GTP to GMP by two successive cleavage steps^{4,5,10}. The crystal structure of full-length hGBP1 (termed hGBP1^{FL} hereafter) has been solved in the nucleotide-free state and in complex with the nonhydrolysable GTP analogue GppNHp^{3,11}. The structures show that hGBP1 has a 36-kDa N-terminal large G (LG) domain and a 32-kDa carboxy-terminal α -helical elongated domain. Instead of the canonical (N/T)KxD motif, hGBP1 uses an RD motif to bind the guanine base¹¹. Although the structure of the dynamin G domain itself is different¹², hGBP1 is considered to belong to the dynamin family of GTP-binding proteins owing to similar domain organization and biochemical properties. The catalytic machinery and the contribution of the different domains to the GTPase reaction remain elusive, and conflicting models have been presented for the role of the C-terminal GTPase effector domain of dynamin^{13,14}.

Structural analysis of hGBP1^{FL} has not revealed the mechanism of GTP cleavage, because the dimers observed in the crystals did not adopt a catalytically competent conformation. Here we focus our attention on the LG domain (residues 1–317; termed hGBP1^{LG}

hereafter), which retains the main biochemical properties of hGBP1^{FL} such as nucleotide affinity, GTP-induced homodimerization, co-operative GTPase activity, and forming GMP as the main product after hydrolysis (see below and Supplementary Fig. 1 and Tables 1 and 2). In contrast to hGBP1^{FL}, hGBP1^{LG} hydrolyses extraneously added GDP and forms a stable dimeric GMP•AlF_x complex.

We solved the crystal structures of GppNHp-, GDP•AlF₃-, GMP•AlF₄⁻- and GMP-bound hGBP1^{LG} at 2.8, 3.2, 2.2 and 2.9 Å, respectively. As expected from the solution behaviour, GppNHp-, GDP•AlF₃-, GMP•AlF₄⁻-bound hGBP1^{LG} crystallized as dimers (Fig. 1a). These dimers are very similar and compatible with the presence of the C-terminal helical domains in the model of the full-length proteins (Fig. 1b). They bury a large interface of ~3,900 Å², including residues that are conserved in hGBP isoforms and all regions with nucleotide-dependent conformation (Supplementary Fig. 2). Switches 1 and 2 are thus mostly inaccessible to possible downstream effectors. Mutation of residues involved in dimerization has a considerable impact on nucleotide binding, GTPase activity and product ratio¹⁵. Superimposition of GppNHp-bound dimeric hGBP1^{LG} and monomeric hGBP1^{FL} (ref. 11) shows that the backbone of helix $\alpha 4'$ shifts up to 9 Å towards the previous location of the hGBP1^{FL} C-terminal $\alpha 12$ and $\alpha 13$ helices (Fig. 1c). This shift is mandatory in dimers to avoid steric clash with switch 2 of the dimer counterpart and is not observed in the monomeric GMP-bound hGBP1^{LG} structure. Outside the LG domain, this shift may induce a similar displacement of the $\alpha 12$ and $\alpha 13$ helices that contact helix $\alpha 4'$. In addition, formation of the dimer causes substantial changes in the switch regions. The side chain hydroxyl group of Ser 73 is reoriented to coordinate a nucleophilic water molecule (Wat10) positioned 3 Å away from the γ -phosphate, suggesting that this residue has a catalytic role in the GTP cleavage mechanism (Fig. 2a and Supplementary Fig. 3a). The most striking feature is the flip of Arg 48 of the P-loop into the nucleotide-binding pocket. Its repositioning is apparently induced to avoid steric clashes with the $\beta 4$ and $\alpha 3$ loop, especially with Thr 133 from the other molecule of the dimer. In the monomeric nucleotide-free (not shown) and GMP-bound hGBP1^{LG} structures, this arginine points towards the solvent.

In the structure of the transition-state mimic hGBP1^{LG}•GDP•AlF₃, Arg 48 coordinates AlF₃ and α - and β -phosphate oxygen atoms (Supplementary Fig. 3b). Its position overlaps well with the arginine supplied in *trans* by the GTPase-activating protein RasGAP to the Ras active site¹⁶ (Fig. 2b), suggesting that Arg 48 functions in *cis* to stabilize the transition state and that dimerization (or oligomerization in general) is necessary for full activity. This phenomenon is also seen in other large GTPases such as dynamin² and IIGP^{17,18}, whereas complexes with GAPs are required for Ras-like GTP-binding proteins¹⁹. In the absence of structural knowledge, Arg 48 has been shown to have a crucial role in the GTPase reaction¹⁵ and in

¹Abteilung Strukturelle Biologie, Max-Planck-Institut für Molekulare Physiologie, Otto-Hahn-Straße 11, 44227 Dortmund, Germany. ²Laboratoire d'Enzymologie et Biochimie Structurales, CNRS UPR 9063, Avenue de la Terrasse, 91198 Gif-sur-Yvette, France. [†]Present addresses: Memorial Sloan Kettering Cancer Center, Structural Biology Program, 1275 York Avenue, New York, New York 10021, USA (A.G.); Center for Molecular Medicine of Cologne (CMMC), Institut für Genetik, Zülpicher Straße 47, 50674 Köln, Germany (G.J.K.P.); Physikalische Chemie 1, Ruhr-Universität Bochum, 44780 Bochum, Germany (C.H.).

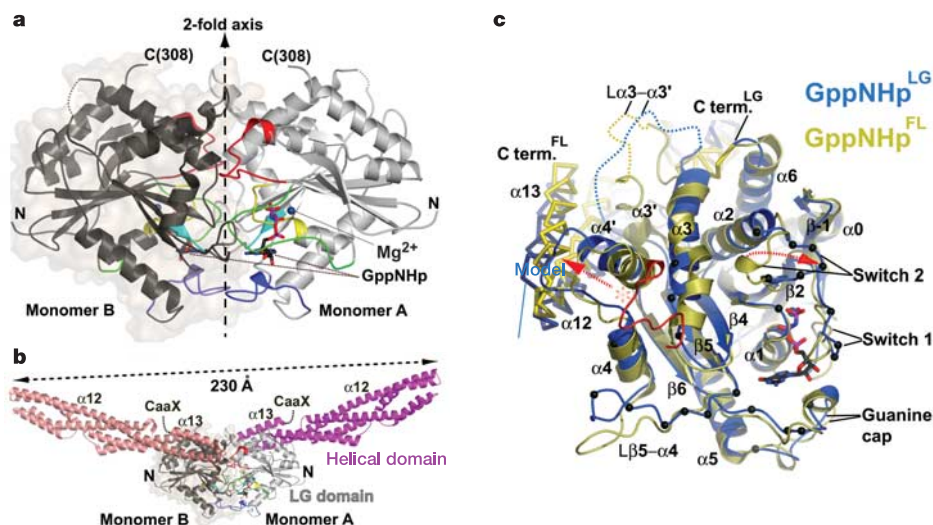


Figure 1 | View of the hGBP1 dimer and its interface. **a**, The two-fold symmetry related hGBP1^{LG}•GppNHp dimer buries most of the regions that have nucleotide-dependent conformation. The P-loop is represented in yellow, switch 1 in green, switch 2 in red, RD₁₈₄ in cyan and the guanine cap in lilac. **b**, Model of the hGBP1^{FL}•GppNHp dimer obtained by superimposing monomeric hGBP1^{FL} on dimeric hGBP1^{LG}. Localization of the prenylation CaaX motifs at hGBP1^{FL} C termini is indicated. **c**, Superimposition of hGBP1^{LG}•GppNHp (blue) and hGBP1^{FL}•GppNHp¹¹

(yellow) structures viewed from the dimer interface. Black spheres indicate residues involved in the interface and red arrows indicate significant displacements between structures. The conformational change to avoid steric clashes (red star) with switch 2 of the dimer counterpart (red ribbon) is thought to displace helix $\alpha 4'$ (red arrow), which in turn would induce changes in the location of the C-terminal helices $\alpha 12$ and $\alpha 13$ (yellow C α trace) of hGBP1^{FL}, which have thus been remodelled (blue C α trace).

localization of hGBP1^{FL} to the Golgi²⁰. Here, we introduced R48A and S73A mutations into both hGBP1^{FL} and hGBP1^{LG}. We found that the mutants are stable in the absence of nucleotide and show similar binding affinity for nucleotides (Supplementary Table 1). In contrast to wild type, the intrinsic GTPase activity of the mutants in the FL and LG domain background does not increase with increasing protein concentration (Fig. 3a).

The GDP cleavage reaction of hGBP1^{LG} to GMP and phosphate is also cooperative (Fig. 3b) and has a catalytic rate constant (k_{cat}) of $\sim 150 \text{ min}^{-1}$ under saturating conditions. The dimerization constant of $9 \mu\text{M}$ is much higher than that of the GTPase reaction ($0.3 \mu\text{M}$; $k_{cat} = 119 \text{ min}^{-1}$). Because hGBP1^{FL} cannot form dimers with GMP•AlF₃, we assume that its dimerization constant in the presence of GDP is too low to allow formation of a productive dimer and cleavage of GDP. For the GTPase reaction, however, we assume that further cleavage of GDP is faster than the dissociation of the dimer, thereby explaining the successive cleavage of both phosphates.

In the GMP•AlF₄⁻-bound dimeric structure, the guanine cap has moved slightly within the interface as compared with the dimeric complexes described above (Fig. 4a). There is a marked change in the torsional angles around the nucleotide ribose moiety when comparing the GMP•AlF₄⁻ and GDP•AlF₃ structures (Supplementary Table 3). The torsional angle of the C4'–C5' bond shows a significant change of 120° , going from -62° to 58° . As a result, the ribose plane of GMP moves by $\sim 2.5 \text{ \AA}$ and the α -phosphate now occupies the canonical position of the β -phosphate with its coordination with the P-loop and Mg²⁺. Aluminium fluoride, modelled as AlF₄⁻, occupies the AlF₃/ γ -phosphate position of the GDP•AlF₃/GppNHp complexes with similar coordinations (Supplementary Fig. 3a–c). To validate the contribution of catalytic residues, we studied the effects of R48A and S73A mutants on the hGBP1^{LG} GDPase activity. Consistent with structural observation, both mutants show a low intrinsic activity and the cooperativity of the reaction is completely abolished (Fig. 3b). This argues in favour of an identical mechanism for both the GTPase and the GDPase activity, attributing crucial roles to Arg 48 and Ser 73.

The structure of the hGBP1^{LG}•GMP product complex represents the monomer observed in solution. The switch and guanine cap

regions relax into a conformation away from the nucleotide. Consequently, the switch regions become partially flexible. Residues ordered in these regions have their backbone shifted between 6 and 9 Å (Supplementary Fig. 4). Switch 2 adopts a conformation that would clash with the second monomer in the dimer structures described above. We conclude that loss of the second phosphate induces conformational changes that, in turn, lead to dissociation of the dimer. Concomitant with reorganization of the nucleotide-binding site, the torsional angle of the C4'–C5' bond of the GMP ribose is increased by a further 56° to 114° as compared to GMP•AlF₄⁻ (Supplementary Table 3), but the position of α -phosphate relative to the P-loop and the base interaction with Asp 184 are unaltered. The guanine cap backbone of Leu 247 and Leu 250 shows a shift of $\sim 8\text{--}9 \text{ \AA}$, thereby exposing the guanine base for possible interaction with Phe 65 from switch 1 (Supplementary Fig. 3d). Altogether, these observations might explain the ability of hGBP1 to bind GMP with

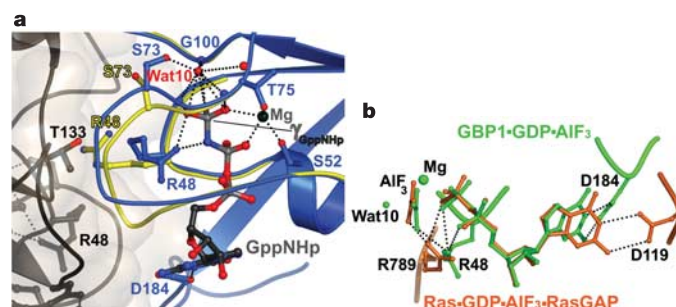


Figure 2 | Structural analysis of the GTPase reaction. **a**, Comparison of the GppNHp•Mg²⁺-binding pockets of hGBP1^{LG} (blue) and hGBP1^{FL} (yellow) highlights the dimerization-induced reorientation of the catalytic Arg 48 and Ser 73 side chains on their corresponding loops. The grey van der Waals surface of monomer B (black) from the hGBP1^{LG}•GppNHp dimer is shown to indicate how Arg 48 of monomer A would clash with Thr 133 from monomer B. **b**, GDP•AlF₃ from the Ras•RasGAP complex¹⁶ (orange) is superimposed on GDP•AlF₃ from hGBP1^{LG} (green), indicating that the *cis* 'arginine finger' of hGBP1 (R48) has a similar orientation to that of the *trans* arginine from RasGAP (R789).

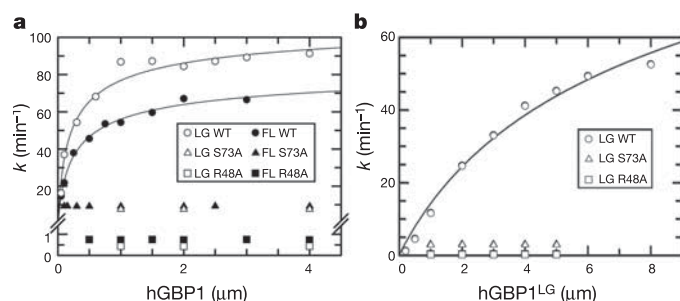


Figure 3 | Biochemical analysis of the catalytic mechanism. The concentration-dependent GTPase activity is used as an assay for cooperativity. **a**, Concentration dependence of the multiturnover-specific GTPase activity gives a k_{cat} of 88 and 110 min^{-1} , and a dimerization constant of 0.4 and 0.3 μM for wild-type hGBP1^{FL} and hGBP1^{LG}, respectively. The hGBP1^{LG} S73A mutant shows a higher basal activity ($\sim 9 \text{ min}^{-1}$) than the hGBP1^{LG} R48A mutant ($\sim 0.7 \text{ min}^{-1}$). These values are similar to those of the hGBP1^{FL} S73A ($\sim 8 \text{ min}^{-1}$) and R48A ($\sim 0.8 \text{ min}^{-1}$) mutants. Note the non-cooperative cleavage by the R48A and S73A mutants. **b**, Concentration dependence of the GDPase activities of wild-type hGBP1^{LG} and the S73A and R48A mutants, measured as in **a**. Both mutants show a low intrinsic activity as compared with wildtype ($\sim 3 \text{ min}^{-1}$ for S73A and $\sim 0.3 \text{ min}^{-1}$ for R48A) and no cooperativity.

affinity similar to that for GDP and GTP. In other GTP-binding proteins such as Ras, which cannot shift the α -phosphate into the P-loop or adopt a nucleotide conformation like that in hGBP1 (Fig. 4b), the affinity for GMP is lower by six orders of magnitude as compared with GDP and GTP²¹.

We have shown that efficient GTP or GDP cleavage by the LG domain of hGBP1 requires dimerization in the presence of guanine nucleotides with two or three phosphates (and their AlF_x mimics; Supplementary Fig. 5). The juxtaposition of nucleotide-binding regions in the dimer interface for reciprocal GTPase stimulation has been reported *in trans* for the pea TOC34 homodimer²² and for the SRP–SR heterodimer, in which the nucleotide itself acts *in trans* and protein residues act *in cis* with the substrate to mediate catalysis^{23,24}. Here, dimerization reorients the catalytic residue Arg 48 into its own active site such that it can participate in catalysis. Although an arginine at the fourth position of the P-loop is absent in dynamin, it is present in Atlastins²⁵ and nucleotide kinases^{26,27}. The other crucial residue for dimerization-induced catalysis in hGBP1 is Ser 73 in switch 1, which reorients itself such that it contacts the nucleophilic water and functions to localize and/or to activate the water for nucleophilic attack. Serine or threonine residues from switch 1 have been proposed to have a similar function in the ATPase reaction of myosins²⁸. In dynamin 1, mutation of a serine residue in switch 1 strongly decreases the rate of assembly-stimulated GTPase²⁹.

Ras-like and other GTP-binding proteins are signalling molecules in which a structural change in the switch regions mediates and terminates interactions with downstream effectors. In hGBP1, the switch regions are involved in dimerization and stay buried until the hydrolysis of GTP to GMP is complete, and are thereby inaccessible for effectors. Homodimerization may induce reciprocal conformational changes towards the C terminus outside the LG domain. It generates an extended dimer of 230 Å length (Fig. 1b) that positions the prenylated C termini on the same side for membrane anchoring, in line with the Golgi targeting of hGBP1 induced by aluminium fluoride treatment in cells²⁰. This orients the extended C-terminal helical domains of the slightly curved hGBP1 dimer (Fig. 1b) towards a similarly curved membrane to perform the hitherto unknown function of hGBP1.

METHODS

Proteins. We cloned hGBP1^{LG} (residues 1–317 of hGBP1) and generated mutants of hGBP1^{LG} or hGBP1^{FL} in the plasmid pQE9 (Qiagen) using standard techniques (see Supplementary Methods).

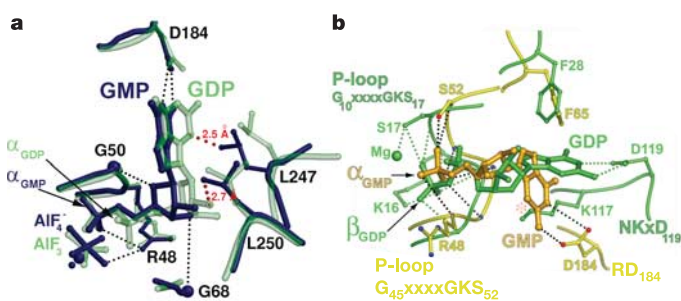


Figure 4 | Structural analysis of the GDPase reaction. **a**, Superimposition of the active sites of hGBP1^{LG}•GMP•AlF₄⁻ (blue) and hGBP1^{LG}•GDP•AlF₃⁻ (green) structures, respectively, showing the shift of GMP for the second hydrolysis step. Black broken lines show stabilizing polar interactions and red broken lines indicate unfavourable vicinities between the nucleotide in the GMP•AlF₄⁻ structure and the guanine cap residues as found in the GMP•AlF₃⁻ structure. **b**, Superimposition of nucleotide-binding sites of hGBP1^{LG}•GMP (yellow, gold) and Ras•GDP (green; Protein Data Bank accession code 4Q21) structures with the α -phosphate of GMP occupying a similar position to that of the β -phosphate of Ras•GDP. Arg 48 is pointing away from the active site. Red star indicates possible steric hindrance between Lys 117 of the (N/T)KxD motif from Ras and the GMP base conformation found in hGBP1^{LG}.

Biochemical analysis. The GTPase and GDPase activities of wild-type and S73A and R48A mutant hGBP1^{FL} and hGBP1^{LG} were measured as described¹⁵. We mixed 300 μM GTP or GDP and varying concentrations of protein at 37 °C in a buffer containing 50 mM Tris–HCl (pH 8.0) and 5 mM MgCl₂. Aliquots were taken at adequate time intervals, and the progress of hydrolysis was analysed by reversed phase chromatography (C-18, Hypersil). Rates were derived from a linear fit to the initial reaction (<30% GTP hydrolysed) and plotted against the protein concentration used. Equilibrium binding constants for the dimerization were obtained using a quadratic binding model¹⁵.

Crystallization and crystallography. All crystallizations were done by vapour diffusion methods at protein concentrations between 300 and 900 μM in 50 mM Tris (pH 8.0), 50 mM NaCl, 3 mM 1,4-dithioerythritol (DTE) and 5–10 mM MgCl₂. Solutions used in crystallizations were at neutral pH except as stated below. Crystals of the hGBP1^{LG}•GppNHp•Mg²⁺ complex appeared overnight in 10% polyethylene glycol 3350 (PEG3350), 100 mM sodium thiocyanate and 40 mM CaCl₂ with 10 mM GppNHp and 10 mM MgCl₂. Crystals of the hGBP1^{LG}•GDP•AlF₃•Mg²⁺ complex appeared within a week in 12.5% PEG3350, 125 mM Na₂HPO₄ (PEG/ion screen supplied at pH 9.1, Hampton Research) with 15 mM GDP, 10 mM MgCl₂, 5 mM AlCl₃, 60 mM NaF and 2 mM DTE. Crystals of the hGBP1^{LG}•GMP•AlF₄⁻•Mg²⁺ complex appeared within a week in 5% PEG3350, 150 mM CaCl₂ with 20 mM GMP, 9 mM MgCl₂, 5 mM AlCl₃, 60 mM NaF and 3 mM DTE. Crystals of hGBP1^{LG}•GMP appeared within a week in 20% PEG3350 and 130 mM NH₄NO₃ with 15 mM GMP.

Data collections were done at 100 K at the European Synchrotron Radiation Facility (ESRF), as summarized in Supplementary Table 4. Initial phases were obtained by molecular replacement with AmoRe or MOLREP in the CCP4 suite³⁰. The GppNHp-bound hGBP1^{LG} crystals contain one molecule in the asymmetric unit, whereas the GDP•AlF₃, GMP•AlF₄⁻ and GMP-bound complexes contain two molecules in the asymmetric unit. The GMP-bound hGBP1^{LG} dimer observed in the crystal asymmetric unit is not compatible with the presence of the hGBP1^{FL} C-terminal helical domain and corresponds to a nonrelevant crystal dimer burying a smaller interface (1,200 Å²) than found in the GDP•AlF₃ and GMP•AlF₄⁻ dimers. No electron density for Mg²⁺ is found in the active site of the GMP-bound complex, consistent with the notion that GMP binding is Mg²⁺ independent (data not shown). Details regarding the final model are given in Supplementary Methods. Supplementary Fig. 6 shows examples of electron-density simulated-annealing omit maps for all crystal structures.

Received 8 August; accepted 30 November 2005.

1. Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H. & Schreiber, R. D. How cells respond to interferons. *Annu. Rev. Biochem.* **67**, 227–264 (1998).
2. Praefcke, G. J. & McMahon, H. T. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nature Rev. Mol. Cell Biol.* **5**, 133–147 (2004).
3. Prakash, B., Praefcke, G. J., Renault, L., Wittinghofer, A. & Herrmann, C. Structure of human guanylate-binding protein 1 representing a unique class of GTP-binding proteins. *Nature* **403**, 567–571 (2000).

4. Schwemmle, M. & Staeheli, P. The interferon-induced 67-kDa guanylate-binding protein (hGBP1) is a GTPase that converts GTP to GMP. *J. Biol. Chem.* **269**, 11299–11305 (1994).
5. Praefcke, G. J., Geyer, M., Schwemmle, M., Robert Kalbitzer, H. & Herrmann, C. Nucleotide-binding characteristics of human guanylate-binding protein 1 (hGBP1) and identification of the third GTP-binding motif. *J. Mol. Biol.* **292**, 321–332 (1999).
6. Cheng, Y. S., Patterson, C. E. & Staeheli, P. Interferon-induced guanylate-binding proteins lack an N(T)KXD consensus motif and bind GMP in addition to GDP and GTP. *Mol. Cell. Biol.* **11**, 4717–4725 (1991).
7. Anderson, S. L., Carton, J. M., Lou, J., Xing, L. & Rubin, B. Y. Interferon-induced guanylate binding protein-1 (GBP-1) mediates an antiviral effect against vesicular stomatitis virus and encephalomyocarditis virus. *Virology* **256**, 8–14 (1999).
8. Guenzi, E. *et al.* The helical domain of GBP-1 mediates the inhibition of endothelial cell proliferation by inflammatory cytokines. *EMBO J.* **20**, 5568–5577 (2001).
9. Guenzi, E. *et al.* The guanylate binding protein-1 GTPase controls the invasive and angiogenic capability of endothelial cells through inhibition of MMP-1 expression. *EMBO J.* **22**, 3772–3782 (2003).
10. Neun, R., Richter, M. F., Staeheli, P. & Schwemmle, M. GTPase properties of the interferon-induced human guanylate-binding protein 2. *FEBS Lett.* **390**, 69–72 (1996).
11. Prakash, B., Renault, L., Praefcke, G. J., Herrmann, C. & Wittinghofer, A. Triphosphate structure of guanylate-binding protein 1 and implications for nucleotide binding and GTPase mechanism. *EMBO J.* **19**, 4555–4564 (2000).
12. Niemann, H. H., Knetsch, M. L., Scherer, A., Manstein, D. J. & Kull, F. J. Crystal structure of a dynamin GTPase domain in both nucleotide-free and GDP-bound forms. *EMBO J.* **20**, 5813–5821 (2001).
13. Sever, S., Muhlberg, A. B. & Schmid, S. L. Impairment of dynamin's GAP domain stimulates receptor-mediated endocytosis. *Nature* **398**, 481–486 (1999).
14. Marks, B. *et al.* GTPase activity of dynamin and resulting conformation change are essential for endocytosis. *Nature* **410**, 231–235 (2001).
15. Praefcke, G. J. *et al.* Identification of residues in the human guanylate-binding protein 1 critical for nucleotide binding and cooperative GTP hydrolysis. *J. Mol. Biol.* **344**, 257–269 (2004).
16. Scheffzek, K. *et al.* The Ras–RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333–338 (1997).
17. Uthaiyah, R. C., Praefcke, G. J., Howard, J. C. & Herrmann, C. IIGP1, an interferon- γ -inducible 47-kDa GTPase of the mouse, showing cooperative enzymatic activity and GTP-dependent multimerization. *J. Biol. Chem.* **278**, 29336–29343 (2003).
18. Ghosh, A., Uthaiyah, R., Howard, J., Herrmann, C. & Wolf, E. Crystal structure of IIGP1: a paradigm for interferon-inducible p47 resistance GTPases. *Mol. Cell* **15**, 727–739 (2004).
19. Scheffzek, K., Ahmadian, M. R. & Wittinghofer, A. GTPase-activating proteins: helping hands to complement an active site. *Trends Biochem. Sci.* **23**, 257–262 (1998).
20. Modiano, N., Lu, Y. E. & Cresswell, P. Golgi targeting of human guanylate-binding protein-1 requires nucleotide binding, isoprenylation, and an IFN- γ -inducible cofactor. *Proc. Natl Acad. Sci. USA* **102**, 8680–8685 (2005).
21. John, J. *et al.* Kinetics of interaction of nucleotides with nucleotide-free H-ras p21. *Biochemistry* **29**, 6058–6065 (1990).
22. Sun, Y. J. *et al.* Crystal structure of pea Toc34, a novel GTPase of the chloroplast protein translocon. *Nature Struct. Biol.* **9**, 95–100 (2002).
23. Focia, P. J., Shepotinovskaya, I. V., Seidler, J. A. & Freymann, D. M. Heterodimeric GTPase core of the SRP targeting complex. *Science* **303**, 373–377 (2004).
24. Egea, P. F. *et al.* Substrate twinning activates the signal recognition particle and its receptor. *Nature* **427**, 215–221 (2004).
25. Zhu, P. P. *et al.* Cellular localization, oligomerization, and membrane association of the hereditary spastic paraplegia 3A (SPG3A) protein atlastin. *J. Biol. Chem.* **278**, 49063–49071 (2003).
26. Lavie, A. *et al.* Crystal structure of yeast thymidylate kinase complexed with the bisubstrate inhibitor P1-(5'-adenosyl) P5-(5'-thymidyl) pentaphosphate (TP5A) at 2.0 Å resolution: implications for catalysis and AZT activation. *Biochemistry* **37**, 3677–3686 (1998).
27. Leippe, D. D., Koonin, E. V. & Aravind, L. Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* **333**, 781–815 (2003).
28. Fisher, A. J. *et al.* X-ray structures of the myosin motor domain of *Dictyostelium discoideum* complexed with MgADP.BeF₃ and MgADP.AIF₄. *Biochemistry* **34**, 8960–8972 (1995).
29. Song, B. D., Leonard, M. & Schmid, S. L. Dynamin GTPase domain mutants that differentially affect GTP binding, GTP hydrolysis, and clathrin-mediated endocytosis. *J. Biol. Chem.* **279**, 40431–40436 (2004).
30. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the ESRF for providing synchrotron radiation facilities and the staff of beamlines BM30A and ID14-EH1/4 for technical assistance during data collection; and E. Wolf for discussions, and M.-F. Carlier and J. Cherfils for support. This work was supported by grants from the Association pour la Recherche contre le Cancer (to L.R.) and the Boehringer Ingelheim Fonds (to G.J.K.P.).

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 2BC9 (GppNHP-bound hGBP1^{LG}), 2B92 (GDP•AlF₃-bound hGBP1^{LG}), 2B8W (GMP•AlF₄-bound hGBP1^{LG}) and 2D4H (GMP-bound hGBP1^{LG}). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.R. (Louis.Renault@lebs.cnrs-gif.fr) or A.W. (alfred.wittinghofer@mpi-dortmund.mpg.de).

Role of genomic instability and p53 in AID-induced *c-myc-Igh* translocations

Almudena R. Ramiro^{1,†}, Mila Jankovic^{1,*}, Elsa Callen³, Simone Difilippantonio³, Hua-Tang Chen³, Kevin M. McBride¹, Thomas R. Eisenreich¹, Junjie Chen⁴, Ross A. Dickins^{5,6}, Scott W. Lowe^{5,6}, Andre Nussenzweig³ & Michel C. Nussenzweig^{1,2}

Chromosomal translocations involving the immunoglobulin switch region are a hallmark feature of B-cell malignancies¹. However, little is known about the molecular mechanism by which primary B cells acquire or guard against these lesions. Here we find that translocations between *c-myc* and the *Igh* locus (*Igh*) are induced in primary B cells within hours of expression of the catalytically active form of activation-induced cytidine deaminase (AID), an enzyme that deaminates cytosine to produce uracil in DNA^{2,3}. Translocation also requires uracil DNA glycosylase (UNG), which removes uracil from DNA to create abasic sites that are then processed to double-strand breaks^{4,5}. The pathway that mediates aberrant joining of *c-myc* and *Igh* differs from intrachromosomal repair during immunoglobulin class switch recombination in that it does not require histone H2AX⁶, p53 binding protein 1 (53BP1)^{7,8} or the non-homologous end-joining protein Ku80⁹. In addition, translocations are inhibited by the tumour suppressors ATM, Nbs1, p19 (Arf) and p53, which is consistent with activation of DNA damage- and oncogenic stress-induced checkpoints during physiological class switching. Finally, we demonstrate that accumulation of AID-dependent, *Igh*-associated chromosomal lesions is not sufficient to enhance *c-myc-Igh* translocations. Our findings reveal a pathway for surveillance and protection against AID-dependent DNA damage, leading to chromosomal translocations.

AID is essential for both class-switch recombination and somatic hypermutation of antibody genes^{2,10}. In addition, it is required for the accumulation of plasmacytoma-associated *c-myc-Igh* translocations in mice^{11–13}, but the role of AID and immunoglobulin class switching in the aetiology of these chromosome fusions has not been determined. To examine the role of AID in *c-myc-Igh* translocations we used retroviral transduction to express AID in *Aid*^{−/−} B cells stimulated with lipopolysaccharide (LPS) and interleukin (IL)-4 *in vitro* (Supplementary Fig. 1a). Under these conditions AID protein was expressed at a tenfold higher level than in wild-type B cells (Supplementary Fig. 1b), and switch recombination and switch junctions were normal (Supplementary Fig. 2). Reciprocal chromosome 12 and 15 *c-myc-Igh* translocations were detected using a polymerase chain reaction (PCR) assay sensitive to 1–2 translocations (Fig. 1 and Supplementary Fig. 1c), and were confirmed by Southern blotting with *c-myc* and *Igh* probes and by sequence analysis (Fig. 1b and Supplementary Fig. 3). Translocations were readily detected in AID-overexpressing cells, and the junctions were similar to those found in B cells expressing physiological levels of AID *in vitro* (see below) and in IL-6 transgenic mice¹² (Supplementary

Fig. 3). In contrast, translocations were absent from uninfected *Aid*^{−/−} B cells (2.6×10^7 cells tested; see below) and *Aid*^{−/−} B cells expressing a catalytically inactive point mutant of AID (AID^{E58Q}; $>1 \times 10^7$ cells tested; Fig. 1b). We conclude that overexpression of catalytically active AID promotes *c-myc-Igh* translocations in B cells stimulated with LPS and IL-4.

To examine the kinetics of the translocation reaction, we infected *Aid*^{−/−} B cells with retroviruses encoding AID-oestrogen receptor (ER) or AID^{E58Q}-ER (Supplementary Fig. 1a); *c-myc-Igh* fusions were assayed after addition of tamoxifen. Switch recombination and *c-myc-Igh* translocations were both detectable as early as 12 h after tamoxifen addition (Supplementary Fig. 4a and data not shown), and by 24 h translocations were found on average once in 2×10^4 cells (Fig. 1c). In contrast, *Aid*^{−/−} B cells infected with AID^{E58Q}-ER failed to show translocations or class switching at any time (Supplementary Fig. 4a and data not shown). We conclude that *c-myc-Igh* translocations are rapidly induced by AID expression *in vitro*.

It has been suggested that AID might promote accumulation of pre-existing *c-myc-Igh* translocations by enhancing the survival of cells bearing these chromosome fusions¹³. To explore this possibility we infected *Aid*^{−/−} B cells with a retrovirus encoding AID-ER or AID^{E58Q}-ER and human CD4 as a marker for infection (Supplementary Fig. 1a), and measured cell division by labelling with carboxy-fluorescein diacetate succinimidyl ester (Supplementary Fig. 4b). *Aid*^{−/−} B cells infected with AID-ER or AID^{E58Q}-ER divided at equal rates (averaging one division after 12 h and four divisions after 48 h) with no difference in the rate of cell death (Supplementary Fig. 4 and data not shown). The frequency of translocations found in *Aid*^{−/−} B cells was less than 4×10^{-8} (see below), and increased by orders of magnitude in 24 h in B cells overexpressing retrovirally encoded AID, by which time the cells had undergone only 2–3 divisions (Supplementary Fig. 4). Therefore, AID could not promote sufficient outgrowth of pre-existing translocation-positive cells and instead produces *c-myc-Igh* translocations *de novo* by a mechanism that requires cytidine deaminase activity.

AID is thought to initiate the switch reaction by deaminating cytidine to produce U·G mismatches in immunoglobulin switch DNA that are processed to double-strand break (DSB) intermediates^{2,3,14}. To determine whether *c-myc-Igh* translocations require base excision repair we assayed *Ung*^{−/−} B cells¹⁵. Occasional faint bands were detectable in *Ung*^{−/−} B cells infected with retroviruses encoding either AID or catalytically inactive AID^{E58Q}, but these failed to hybridize with both *c-myc* and *Igh* probes in Southern blots and thus were not authentic *c-myc-Igh* translocations (Fig. 1b).

¹Laboratory of Molecular Immunology, The Rockefeller University and ²Howard Hughes Medical Institute, New York, New York 10021, USA. ³Experimental Immunology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁴Division of Oncology, Mayo Clinic, Rochester, Minnesota 55905, USA. ⁵Cold Spring Harbor Laboratory and ⁶Howard Hughes Medical Institute, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. †Present Address: Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain.

*These authors contributed equally to this work.

c-myc-Igh translocations were also absent from *Ung*^{-/-} B cells stimulated to express physiological levels of AID *in vitro* upon treatment with LPS and IL-4 (Fig. 2, $P = 0.0097$ versus wild type (see below)). Thus, like normal immunoglobulin class switching and somatic mutation, efficient translocation requires processing of AID-induced U·G mismatches by the base excision repair protein UNG.

DSB intermediates in the switch reaction are repaired by a non-homologous end-joining (NHEJ) pathway requiring Ku80/Ku70 (refs 9, 16) and partially dependent on DNA-PKcs^{17,18}. To determine whether Ku80 is also required for *c-myc-Igh* translocations we overexpressed AID by retroviral transduction in immunoglobulin transgenic *Ku80*^{-/-} B cells stimulated with LPS and IL-4 (*Ku80*^{-/-}HL)⁹. These cells did not undergo switching but showed authentic *c-myc-Igh* translocations (Fig. 1b), which were also independent of DNA-PKcs catalytic activity (data not shown). Thus, Ku80 is dispensable for fusing *c-myc* and *Igh*, suggesting that an alternative non-classical NHEJ pathway may be involved in aberrant interchromosomal joining.

To determine whether physiological levels of AID expression can produce translocations *in vitro* we assayed wild-type and *Aid*^{-/-} B cells before and after stimulation with LPS and IL-4. Although translocations were infrequent in stimulated wild-type B cells they were absent in stimulated *Aid*^{-/-} B cells (Fig. 2, $P = 0.0097$ for wild type versus *Aid*^{-/-}) and in unstimulated wild-type or *Aid*^{-/-} B cells

(Supplementary Fig. 5). Thus, physiological levels of AID can produce rare *c-myc-Igh* translocations in stimulated wild-type B cells, a finding consistent with the detection of these events in B cells isolated from immunized mice¹⁹.

In addition to NHEJ, efficient class switching requires the DSB response proteins ATM^{20,21}, Nbs1 (refs 22, 23), H2AX (refs 6, 24) and 53BP1 (refs 7, 8). Absence of 53BP1 produces the most profound reduction in switching in mice^{7,8} (Fig. 2), whereas milder defects are found in *Atm*^{-/-} (refs 20, 21), *H2ax*^{-/-} (refs 6, 24) and *Nbs1*^{Δ/-} (refs 22, 23) mice (Fig. 2). To determine whether DSBs produced during class switching are channelled into translocations in the absence of efficient repair, we assayed *53BP1*^{-/-}, *Atm*^{-/-}, *H2ax*^{-/-} and *Nbs1*^{Δ/-} (null) B cells. None of the mutant B cells showed evidence of translocation before stimulation with LPS and IL-4 (Supplementary Fig. 5 and data not shown); after stimulation, the frequency of translocation in *53BP1*^{-/-} and *H2ax*^{-/-} B cells was similar to wild type (Fig. 2, $P = 0.427$ and 0.487 versus wild type, respectively). In contrast, *Atm*^{-/-} B cells exhibited significant enhancement in translocation (Fig. 2, $P < 0.0001$ versus wild type) as did *Nbs1*^{Δ/-} B cells, which are defective in ATM activation²⁵ (Fig. 2, $P = 0.0006$ versus wild type). In contrast, Nbs1 hypomorphic mutant B cells (*Nbs1*^{657Δ5}), which are only mildly defective in ATM activation and normal for class switching²⁵, resembled wild-type B cells with respect to translocation (Fig. 2, $P = 0.611$ versus wild type). The finding that a marked reduction in class switching in

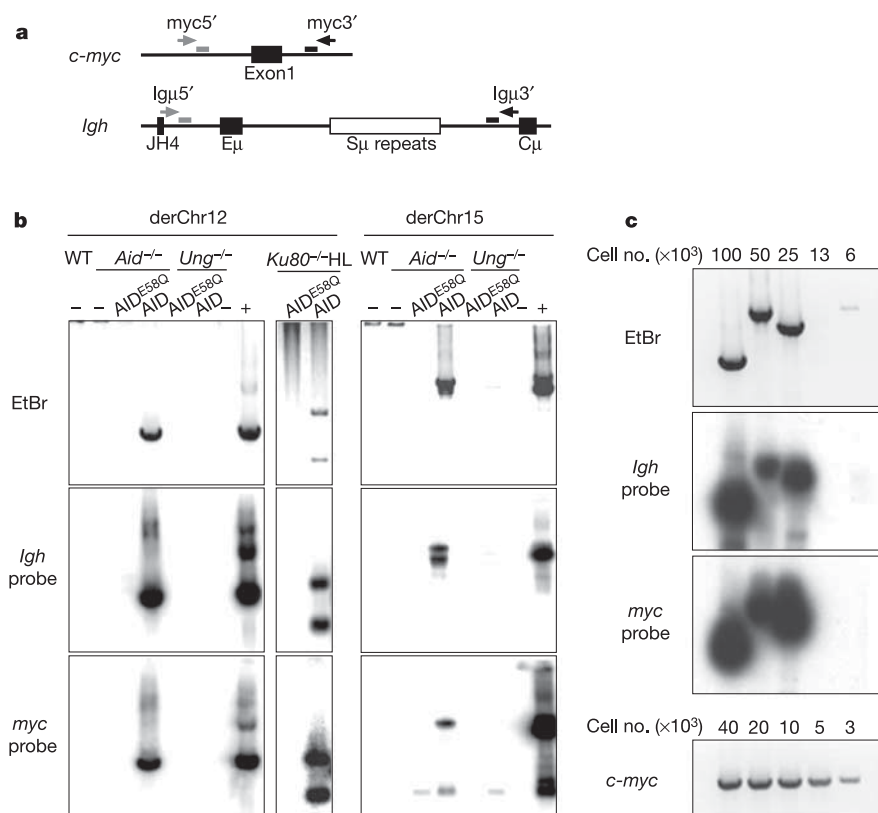


Figure 1 | AID overexpression is sufficient to promote *c-myc-Igh* translocations in B cells. **a**, Schematic representation of the PCR assay used for *c-myc-Igh* translocations. Primers used to detect derivative chromosome 12 (derChr12) and derivative chromosome 15 (derChr15) translocations are represented as horizontal black and grey arrows, respectively. Internal oligonucleotide probes used in Southern blot experiments are shown as horizontal black and grey bars. **b**, AID overexpression promotes *c-myc-Igh* translocations in B cells. Naive B cells from wild type (WT), *Aid*^{-/-}, *Ung*^{-/-} or *Ku80*^{-/-}HL⁹ mice were stimulated with LPS and IL-4 and retrovirally transduced (*Aid*^{-/-}, *Ung*^{-/-}, *Ku80*^{-/-}HL) with AID or E58Q mutant AID (AID^{E58Q}) (Supplementary Fig. 1a) or left uninfected

(WT and *Aid*^{-/-}), as indicated. Representative agarose gels (upper gels) and Southern blots with *Igh* or *myc* oligonucleotide probes (middle and lower gels, respectively) are shown. **c**, *c-myc-Igh* translocations can be detected after 24 h of AID expression. *Aid*^{-/-} B cells were transduced with retroviral vectors encoding AID or catalytically inactive AID^{E58Q} fused to oestrogen receptor (AID-ER or AID^{E58Q}-ER) (Supplementary Fig. 1a). GFP⁺ cells were sorted 24 h after addition of tamoxifen and translocations were analysed by PCR. Agarose gels and Southern blots are shown. Numbers above lanes indicate input cell number (×10³). The lower gel shows *c-myc* PCR as DNA loading control.

53BP1^{-/-} mice did not alter the frequency of translocation, whereas a modest defect in switching in *Atm*^{-/-} mice produced a strong enhancement, indicates that DSB response proteins have distinct roles in the class-switch recombination and translocation reactions.

ATM activates p53 in response to DSBs. To determine whether DNA damage response pathways downstream of ATM are involved in protecting cells from *c-myc-Igh* translocations we examined p53-deficient mice. Whereas freshly isolated *p53*^{-/-} B cells did not carry *c-myc-Igh* translocations (Supplementary Fig. 5), stimulation

with LPS and IL-4 induced a marked increase in the frequency of translocations relative to wild type; this was not due to an increase in AID protein expression in *p53*^{-/-} B cells compared to wild type (Fig. 2, $P < 0.0001$ versus wild type; see also Supplementary Fig. 1b). Notably, loss of a single *p53* allele was sufficient to enhance translocations (Fig. 2, $P = 0.004$ versus wild type). The *c-myc-Igh* translocations in *p53*^{-/-} B cells were similar to translocations found in wild-type B cells (Supplementary Fig. 3) and these events were AID-dependent because B cells deficient in AID and p53 (*Aid*^{-/-}*p53*^{-/-}) produced no *c-myc-Igh* translocations (Fig. 2, $P < 0.0001$ *Aid*^{-/-}*p53*^{-/-} versus *p53*^{-/-}). Thus, like ATM, p53 is essential for protecting B cells from *c-myc-Igh* translocations, but in contrast to ATM, p53 has no detectable effect on the switch reaction (Fig. 2).

p53 is also activated in response to abnormal mitogenic signals, such as those occurring when *c-myc* is deregulated by genomic rearrangements²⁶. Because this form of oncogenic stress induces expression of the p19 (Arf) tumour suppressor, we explored the possibility that p19-mediated activation of p53 might prevent accumulation of *c-myc-Igh* translocations. Although there was no immunoglobulin switch recombination defect in *p19*^{-/-} B cells, the frequency of *c-myc-Igh* translocations was significantly elevated

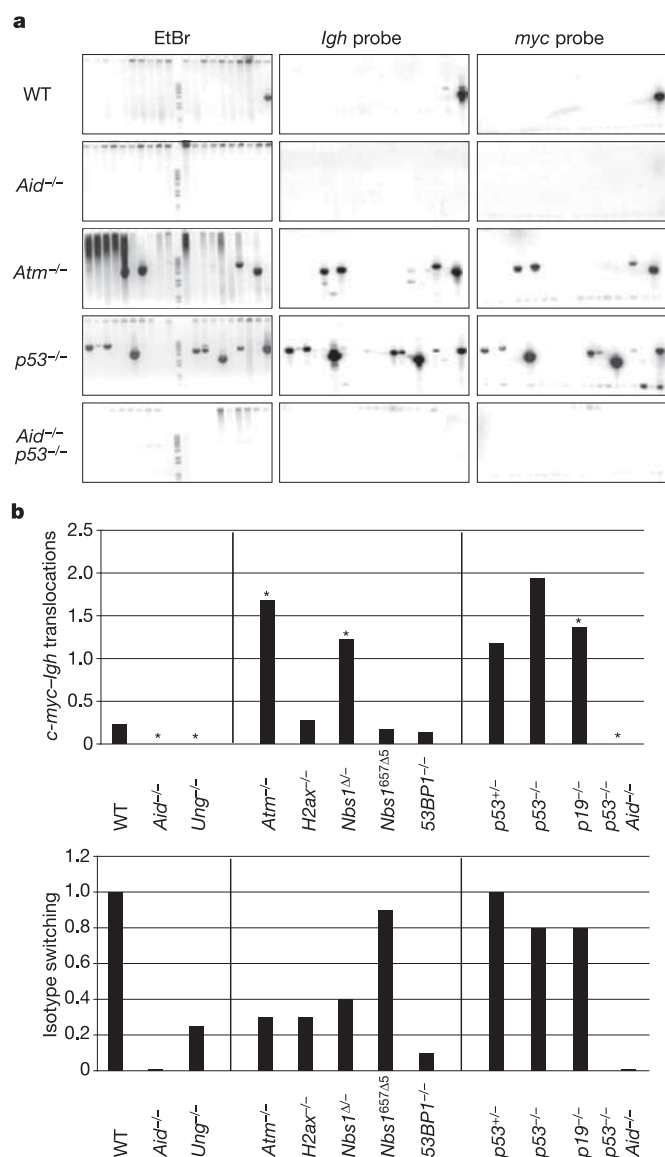


Figure 2 | *c-myc-Igh* translocations in mutant B cells. **a**, *c-myc-Igh* translocations. B cells were cultured with LPS and IL-4 and assayed by PCR after 72 h. Representative agarose gels (left) and Southern blots with *Igh* (middle) and *myc* (right) probes are shown. **b**, Translocation frequency and isotype switching. The upper graph shows the number of *c-myc-Igh* translocations per 10⁷ cells: *Aid*^{-/-} ($P = 0.0097$ versus wild type), *Ung*^{-/-} ($P = 0.0097$ versus wild type), *Atm*^{-/-} ($P < 0.0001$ versus wild type), *H2ax*^{-/-} ($P = 0.487$ versus wild type), *Nbs1*^{Δ/-} ($P < 0.0001$ versus wild type), *Nbs1*^{657Δ5} (hypomorph) ($P = 0.611$ versus wild type), *53BP1*^{-/-} ($P = 0.427$ versus wild type), *p53*^{-/-} ($P < 0.0001$ versus wild type), *p53*^{+/-} ($P = 0.004$ versus wild type), *p19*^{-/-} ($P < 0.0001$ versus wild type) and *p53*^{-/-} *Aid*^{-/-} ($P < 0.0001$ versus *p53*^{-/-}, $P = 0.0152$ versus wild type, $P = 0.611$ versus *Aid*^{-/-}) mutants. Samples with significant P -values versus the wild-type value are labelled with an asterisk. The lower graph shows the efficiency of isotype switching to IgG1 relative to wild type.

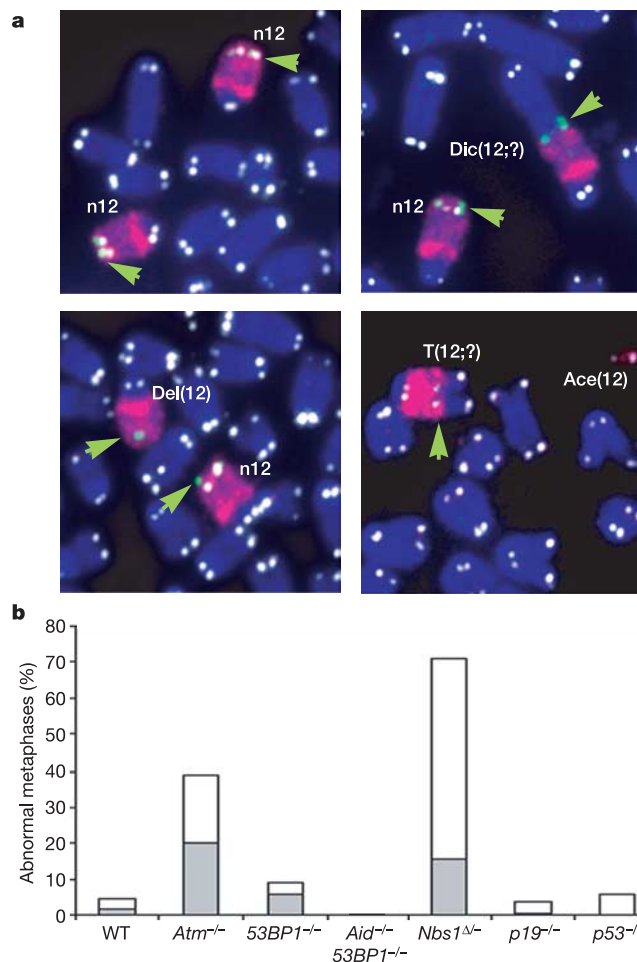


Figure 3 | Chromosomal instability. Metaphase spreads were prepared at 72 h after culture with LPS and IL-4 and hybridized with a combination of painting probes for chromosome 12 (pink), IgH Cα (green) and telomere-specific PNA probes (white), and counterstained with DAPI (blue). **a**, Examples of chromosome-12 associated lesions in mutant B cells include dicentric (Dic), translocations (T), deletions (Del) and acentric fragments (Ace). Normal 12 (n12) is indicated. **b**, Percentage of metaphases with abnormalities specifically associated with chromosome 12 (filled bar, including chromatid and chromosome breaks and fusions) and with all other chromosomes (open bar).

relative to wild type after stimulation with LPS and IL-4 (Fig. 2, $P < 0.0001$ versus wild type). These data suggest that induction of p19 expression is also involved in protection against AID-induced translocations.

To determine whether susceptibility to *c-myc-Igh* translocations correlates with the presence of chromosomal lesions at the IgH locus, we examined metaphase spreads from *53BP1*^{-/-}, *p53*^{-/-}, *p19*^{-/-}, *Atm*^{-/-} and *Nbs1*^{Δ5} B cells that were stimulated with LPS and IL-4 (Fig. 3a). Metaphases were hybridized with a combination of painting probes for chromosome 12 (which carries the mouse IgH locus in a telomere proximal position), IgH Cα (downstream of IgSγ1) and telomere repeats²⁷. Consistent with the proposed defect in synapsis during class switching in *53BP1*^{-/-} B cells^{7,8}, and the observation that 53BP1 accumulates on the IgH locus in cells undergoing switching²⁰, most of the aberrations in *53BP1*^{-/-} metaphases were associated with chromosome 12 (Fig. 3b and Supplementary Table 1). Furthermore, the chromosome 12 lesions in *53BP1*^{-/-} metaphases were AID-dependent because they were absent in *Aid*^{-/-} *53BP1*^{-/-} B cells (Fig. 3b). B cells from *Atm*^{-/-} mice resembled *53BP1*^{-/-} B cells but accumulated a greater number of chromosome lesions (Fig. 3b and Supplementary Table 1). Overall, 39% of the *Atm*^{-/-} metaphases exhibited abnormalities and 20% of aberrations carried either a deletion (24 out of 160) or translocation (6 out of 160) involving chromosome 12. Nine of the metaphases with chromosome 12 deletions lacked a telomere while maintaining a signal for IgH Cα, as would be expected of an unresolved class-switch associated break, whereas 15 cells had also lost IgH Cα (Fig. 4b and Supplementary Table 1). Similarly, a significant fraction of *Nbs1*^{Δ5} B cells exhibited IgH-associated lesions, but these cells also showed a much higher level of general genomic instability (Fig. 3b and Supplementary Table 1). In contrast, loss of p53, which produced no defect in class switching but a strong enhancement of *c-myc-Igh* translocations (Fig. 2), resulted in chromosomal aberrations in 5% of the metaphases analysed but none involving the IgH locus (Fig. 3b and Supplementary Table 1). We conclude that genomic instability is not sufficient to enhance *c-myc-Igh* translocations: similar numbers of translocations were found in p53 and ATM mutant B cells (Fig. 2) that harbour widely differing degrees of genomic instability, whereas *53BP1*^{-/-} cells exhibit AID-dependent IgH instability without a concomitant increase in translocations.

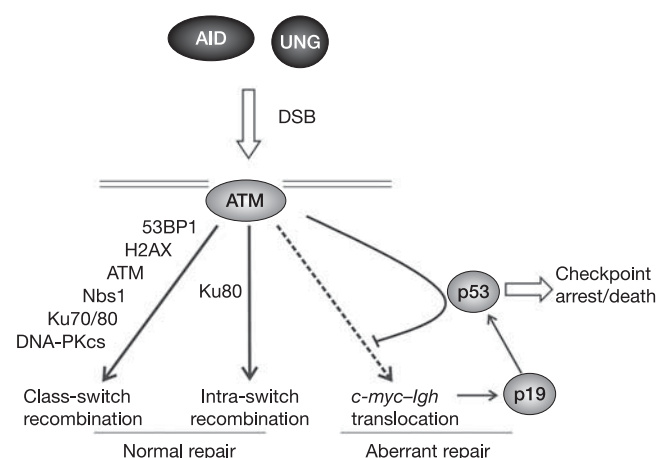


Figure 4 | Distinct pathways mediate class switching and chromosome translocations. AID/UNG-induced DSBs at the IgH locus are normally resolved by 53BP1/Nbs1/ATM/γ-H2AX and NHEJ (Ku70, Ku80 and DNA-PKcs), leading to class switching or intra-switch recombination, which frequently leads to internal Sμ deletions (53BP1/Nbs1/ATM/γ-H2AX-independent and Ku80-dependent)²⁴. Alternatively, DSBs may be resolved by translocations, which are suppressed by ATM- and/or p19-dependent activation of p53.

Our experiments reveal that divergent pathways lead to class switching and *c-myc-Igh* translocation. Both reactions require cytidine deamination and base excision repair and proceed through DSB intermediates, but the factors required for resolution of DNA breaks differ in the two reactions. We would like to propose a working model to account for these findings (Fig. 4). In this scheme, switching requires formation of AID-dependent DSBs, synapsis of upstream switch Sμ with downstream switch breaks mediated by H2AX/53BP1/ATM/Nbs1 (refs 6–8, 20–24) and NHEJ (Ku70/Ku80 and DNA-PKcs)^{9,16–18}. DNA lesions that remain unresolved initiate a p53-dependent checkpoint through the activation of ATM. Cells that escape this initial checkpoint make Ku80-independent translocations that activate p19 in response to deregulation of *c-myc*, thereby triggering p53 by an alternative mechanism (Fig. 4). In this way, the p19–p53 or ATM–p53 signalling axes may have complementary roles in the elimination of incipient cancer cells during different stages of the translocation reaction, while not being needed for class switching *per se*. The model predicts that combined loss of p53 and factors that facilitate immunoglobulin switch synapsis or NHEJ would lead to the accelerated appearance of mature B-cell lymphomas. Consistent with this idea, combined deficiency in H2AX or 53BP1 and p53 leads to rapid lymphoma development (of T- and B-cell lineages), and the mature B-cell lymphomas harbour recurrent *c-myc-Igh* translocations^{28,29}.

Mature B-cell lymphomas are the most common lymphoid malignancies in humans and many of these are associated with cytogenetic abnormalities involving immunoglobulin loci¹. Our data suggest that mutation or decreased expression of p53 contributes early in the pathogenesis of lymphoma by facilitating AID-induced translocations.

METHODS

Retroviral constructs. The AIDER-CD4-IRES construct (Supplementary Fig. 1a, lower panel) was created by replacing an *EcoRI*–*EcoRV* fragment containing the IRES-PuroR cassette of pQCXIP vector (Clontech) with a *BamHI* CD4-IRES fragment obtained by PCR from pMAC54-IRESII plasmid (Miltenyi Biotech) with primers 5′-CGGATCCCCGCCCTCTCCCTCCCCCCCCCTA-3′ and 5′-CGGATCCTCAGTGCCGGCACCTGACACAGAAGA-3′. AIDER or EQER were cloned in *NotI*–*PacI* sites after amplification with primers 5′-GCGG CCGCCGCCACCATGGACAGCCTTCTGATGAAGCA-3′ and 5′-GTTAATTAA TCAGACTGTGGCAGGGAACCCCTCTGC-3′.

PCR and statistical analysis. PCR reactions were performed exactly as described¹². Briefly, genomic DNA corresponding to 10⁵ cells (unless indicated otherwise) was subjected to two rounds of 25 cycles each of nested PCR using the Expand Long Template PCR system (Roche) with the following primers³⁰: for derivative chromosome 12 translocations, first round 5′-TGAGGACCAGAG AGGGATAAAGAGAA-3′ and 5′-GGGGAGGGGGTGTCAAATAAGA-3′, second round 5′-CACCCTGCTATTTCCTTGTGTCTAC-3′ and 5′-GACACCT CCCTTCTACACTCTAAACCG-3′; for derivative chromosome 15 translocations, first round 5′-ACTATGCTATGGACTACTGGGGTCAAG-3′ and 5′-GTGAAAA CCGACTGTGGCCCTGGAA-3′, second round 5′-CCTCAGTCACCGTCT CCTCAGGTA-3′ and 5′-GTGGAGGTGTATGGGGTGTAGAC-3′. For loading control PCR we used the *myc* primers 5′-GGGGAGGGGGTGTCAAATAAT AAGA-3′ and 5′-GTGAAAACCGACTGTGGCCCTGGAA-3′. In Fig. 2b, 6–26 million cells were assayed each from three independent mice: wild type, *Ung*^{-/-}, *Aid*^{-/-} and *p53*^{-/-} *Aid*^{-/-}, 2.6×10^7 ; *Atm*^{-/-}, 2.2×10^7 ; *p53*^{-/-}, 1.6×10^7 ; *53BP1*^{-/-}, *H2ax*^{-/-} and *p19*^{-/-}, 1.4×10^7 ; *Nbs1*^{Δ5}, 0.9×10^7 ; *p53*^{+/-} and *Nbs1*^{657Δ5}, 0.6×10^7 . PCR products were cloned in a TOPO-TA vector (Invitrogen) and sequenced. For Southern blotting DNA was transferred to membranes (GeneScreen, Perkin-Elmer) by alkaline transfer and hybridized to ³²P-end-labelled oligo probes using QuikHyb (Stratagene) according to the manufacturer's instructions. Oligonucleotide probes were internal to the primers used in the PCR assay: derChr15 *c-myc* 5′-GGACTGCGCAGGGAGACCTACAGGGG-3′; derChr15 *Igh* 5′-GAGGGAGCCGGCTGAGAGAAGTTGGG-3′; derChr12 *c-myc* 5′-GCAGCGATTACAGCACTGGGTGCAGG-3′; derChr12 *Igh* 5′-CCT GGTATACAGGACGAACTGCAGCAG-3′. *P*-values for pairwise comparison of translocation frequencies were calculated using one-sided exact Fisher's test (GraphPad software) from contingency tables built with numbers of positive and negative PCR reactions ($n > 150$ for every test).

Mice, B-cell cultures and retroviral transductions, flow cytometry and

cell sorting, and FISH detection of IgH-associated lesions are described in Supplementary Methods.

Received 19 August; accepted 6 December 2005.

Published online 8 January 2006.

1. Kuppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20**, 5580–5594 (2001).
2. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
3. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
4. Rada, C. *et al.* Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. *Curr. Biol.* **12**, 1748–1755 (2002).
5. Schrader, C. E., Linehan, E. K., Mochegova, S. N., Woodland, R. T. & Stavnezer, J. Inducible DNA breaks in IgS regions are dependent on AID and UNG. *J. Exp. Med.* **202**, 561–568 (2005).
6. Petersen, S. *et al.* AID is required to initiate Nbs1/γ-H2AX focus formation and mutations at sites of class switching. *Nature* **414**, 660–665 (2001).
7. Ward, I. M. *et al.* 53BP1 is required for class switch recombination. *J. Cell Biol.* **165**, 459–464 (2004).
8. Manis, J. P. *et al.* 53BP1 links DNA damage-response pathways to immunoglobulin heavy chain class-switch recombination. *Nature Immunol.* **5**, 481–487 (2004).
9. Casellas, R. *et al.* Ku80 is required for immunoglobulin isotype switching. *EMBO J.* **17**, 2404–2411 (1998).
10. Revy, P. *et al.* Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565–575 (2000).
11. Potter, M. & Wiener, F. Plasmacytomagenesis in mice: model of neoplastic development dependent upon chromosomal translocations. *Carcinogenesis* **13**, 1681–1697 (1992).
12. Ramiro, A. R. *et al.* AID is required for *c-myc*/IgH chromosome translocations *in vivo*. *Cell* **118**, 431–438 (2004).
13. Unniraman, S., Zhou, S. & Schatz, D. G. Identification of an AID-independent pathway for chromosomal translocations between the *IgH* switch region and *Myc*. *Nature Immunol.* **5**, 1117–1123 (2004).
14. Imai, K. *et al.* Human uracil-DNA glycosylase deficiency associated with profoundly impaired immunoglobulin class-switch recombination. *Nature Immunol.* **4**, 1023–1028 (2003).
15. Endres, M. *et al.* Increased postischemic brain injury in mice deficient in uracil-DNA glycosylase. *J. Clin. Invest.* **113**, 1711–1721 (2004).
16. Manis, J. P. *et al.* Ku70 is required for late B cell development and immunoglobulin heavy chain class switching. *J. Exp. Med.* **187**, 2081–2089 (1998).
17. Manis, J. P., Dudley, D., Kaylor, L. & Alt, F. W. IgH class switch recombination to IgG1 in DNA-PKcs-deficient B cells. *Immunity* **16**, 607–617 (2002).
18. Bosma, G. C. *et al.* DNA-dependent protein kinase activity is not required for immunoglobulin class switching. *J. Exp. Med.* **196**, 1483–1495 (2002).
19. Roschke, V., Kopantzev, E., Dertzbaugh, M. & Rudikoff, S. Chromosomal translocations deregulating *c-myc* are associated with normal immune responses. *Oncogene* **14**, 3011–3016 (1997).
20. Reina-San-Martin, B., Chen, H. T., Nussenzweig, A. & Nussenzweig, M. C. ATM is required for efficient recombination between immunoglobulin switch regions. *J. Exp. Med.* **200**, 1103–1110 (2004).
21. Lumsden, J. M. *et al.* Immunoglobulin class switch recombination is impaired in *Atm*-deficient mice. *J. Exp. Med.* **200**, 1111–1121 (2004).
22. Reina-San-Martin, B., Nussenzweig, M. C., Nussenzweig, A. & Difilippantonio, S. Genomic instability, endoreduplication, and diminished Ig class-switch recombination in B cells lacking Nbs1. *Proc. Natl Acad. Sci. USA* **102**, 1590–1595 (2005).
23. Kracker, S. *et al.* Nibrin functions in Ig class-switch recombination. *Proc. Natl Acad. Sci. USA* **102**, 1584–1589 (2005).
24. Reina-San-Martin, B. *et al.* H2AX is required for recombination between immunoglobulin switch regions but not for intra-switch region recombination or somatic hypermutation. *J. Exp. Med.* **197**, 1767–1778 (2003).
25. Difilippantonio, S. *et al.* Role of Nbs1 in the activation of the Atm kinase revealed in humanized mouse models. *Nature Cell Biol.* **7**, 675–685 (2005).
26. Lowe, S. W. & Sherr, C. J. Tumor suppression by Ink4a-Arf: progress and puzzles. *Curr. Opin. Genet. Dev.* **13**, 77–83 (2003).
27. Difilippantonio, M. J. *et al.* Evidence for replicative repair of DNA double-strand breaks leading to oncogenic translocation and gene amplification. *J. Exp. Med.* **196**, 469–480 (2002).
28. Ward, I. M. *et al.* 53BP1 cooperates with p53 and functions as a haploinsufficient tumour suppressor in mice. *Mol. Cell Biol.* **25**, 10079–10086 (2005).
29. Bassing, C. H. *et al.* Histone H2AX: a dosage-dependent suppressor of oncogenic translocations and tumors. *Cell* **114**, 359–370 (2003).
30. Kovalchuk, A. L. *et al.* IL-6 transgenic mouse model for extraosseous plasmacytoma. *Proc. Natl Acad. Sci. USA* **99**, 1509–1514 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Honjo for *Aid*^{−/−} mice, M. Bosma for *DNA-PKcs*^{−/−} mice, R. Jaenisch for *Ung*^{−/−} mice, A. Singer and E. Besmer for suggestions, K. Velinon for flow cytometry, and L. Stapelton for painting probes. A.N. was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research and a grant from the AT Childrens Project. S.W.L. was supported by grants from the National Cancer Institute. A.R.R. is a Ramon y Cajal investigator from Ministerio de Educacion y Ciencia, Spain. M.C.N. was supported by grants from the NIH and the Leukemia Society. M.C.N. is a Howard Hughes Institute Investigator.

Author Contributions The last two authors (A.N. and M.C.N.) contributed equally to this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.N. (andre_nussenzweig@nih.gov) or M.C.N. (nussen@mail.rockefeller.edu).

LETTERS

An acidic protein aligns magnetosomes along a filamentous structure in magnetotactic bacteria

André Scheffel¹, Manuela Gruska², Damien Faivre¹, Alexandros Linaroudis², Jürgen M. Plitzko² & Dirk Schüler¹

Magnetotactic bacteria are widespread aquatic microorganisms that use unique intracellular organelles to navigate along the Earth's magnetic field. These organelles, called magnetosomes, consist of membrane-enclosed magnetite crystals that are thought to help to direct bacterial swimming towards growth-favouring microoxic zones at the bottom of natural waters¹. Questions in the study of magnetosome formation include understanding the factors governing the size and redox-controlled synthesis of the nano-sized magnetosomes and their assembly into a regular chain in order to achieve the maximum possible magnetic moment, against the physical tendency of magnetosome agglomeration. A deeper understanding of these mechanisms is expected from studying the genes present in the identified chromosomal 'magnetosome island', for which the connection with magnetosome synthesis has become evident². Here we use gene deletion in *Magnetospirillum gryphiswaldense* to show that magnetosome alignment is coupled to the presence of the *mamJ* gene product. MamJ is an acidic protein associated with a novel filamentous structure, as revealed by fluorescence microscopy and cryo-electron tomography. We suggest a mechanism in which MamJ interacts with the magnetosome surface as well as with a cytoskeleton-like structure. According to our hypothesis, magnetosome architecture represents one of the highest structural levels achieved in prokaryotic cells.

Magnetotactic bacteria are of considerable biophysical interest because they have evolved cellular mechanisms that allow them to exactly meet the requirements for effective magnetic orientation^{1,3}. First, they must precisely control the biomineralization of a magnetite crystal within the single-magnetic-domain size range. Second, growing particles must be properly assembled into a linear chain⁴. In such a configuration, the total magnetic dipole moment is the sum of the moments of individual particles—this generates the maximum possible magnetic moment for magnetotaxis⁵. However, a string of magnetic dipoles has an inherent tendency to agglomerate in order to lower its magnetostatic energy⁶, unless it is adequately stabilized by an organic structure. In addition, such a structure would have to properly anchor the chain, in order to rotate the whole cell into alignment with magnetic field lines⁷.

The mechanisms governing the formation of magnetosome chains, and any structures preventing them from collapse, have remained elusive. Consistent with a tightly regulated biosynthetic assembly, a large genomic 'magnetosome island' was identified as being essential for magnetosome formation in *M. gryphiswaldense*². This magnetosome island encodes a number of specific magnetosome-associated polypeptides, the exact functions of which are unknown. Among them, we identified the conspicuous 426-amino-acid MamJ protein⁸, which is characterized by a high content of acidic amino acids and a repetitive domain structure

(Supplementary Fig. 1), features that have frequently been found in other proteins associated with biomineralization processes⁹.

To verify the assumed essential role of MamJ in the formation of functional magnetosomes, we generated a MamJ-deficient mutant strain ($\Delta mamJ$) (Supplementary Fig. 2). Contrary to our expectations, the mutant was not affected in biomineralization, but produced magnetite crystals identical in size, morphology and

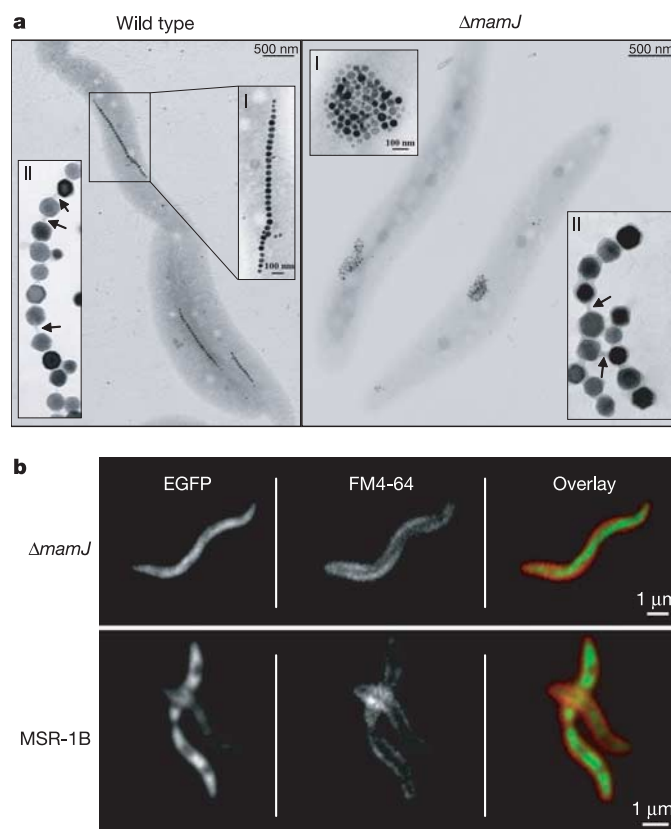


Figure 1 | $\Delta mamJ$ mutant phenotype and intracellular localization of MamJ. **a**, Transmission electron micrographs of wild-type and $\Delta mamJ$ cells. Insets show (I) magnification of magnetosome organization and (II) magnetosome membranes forming junctions (arrows) between isolated crystals. **b**, A MamJ–EGFP fusion protein that restores chain formation in $\Delta mamJ$ cells localizes as a line along the concave cell boundary, but shows punctuate cytoplasmic localization in the non-magnetic deletion mutant MSR-1B, which lacks the magnetosome genes. Cell membranes were stained with the membrane-specific fluorescent dye FM4-64.

¹Max Planck Institute for Marine Microbiology, Celsiusstr. 1, D-28359 Bremen, Germany. ²Department of Structural Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany.

number to wild-type magnetosomes. However, although mutant cells responded to external magnetic fields if examined by microscopy, they showed substantially reduced magnetic orientation in a quantitative light-scattering assay¹⁰, as measured by C_{mag} (the ratio of maximum and minimum scattering intensities; C_{mag} wild type, 1.7; ΔmamJ , 0.5). Notably, transmission electron microscopy (TEM) revealed that the mutant no longer produced straight magnetosome chains, but that magnetite crystals were instead arranged in compact three-dimensional clusters (Fig. 1a).

To determine whether clustering resulted from a loss of magnetosome membrane integrity caused by the ΔmamJ mutation, we isolated magnetosomes from wild-type and mutant cells. Magnetite crystals from ΔmamJ cells were membrane-enclosed and no longer organized in clusters, but spontaneously re-formed in-plane chains or flux-closed rings identical to those of wild-type cells^{8,11};

however, removal of organic material by SDS caused agglomeration of particles. At closer inspection, both wild-type and mutant magnetosomes were tightly interconnected by organic material, which appeared to form junctions between individual particles (Fig. 1a, inset II). These observations indicate that the presence of the magnetosome membrane is required for coherence within the chain, but that the ability to form chains *in vitro* is not affected by loss of MamJ.

We next constructed enhanced green fluorescent protein (EGFP) fusion proteins in order to visualize the subcellular localization of MamJ. *In vivo*, a functional MamJ–EGFP fusion protein was visualized as a linear structure stretching from pole to pole in both wild-type and ΔmamJ cells (Fig. 1b). Consistent with the position of magnetosome chains as revealed by TEM (Fig. 1a), the fluorescent structure is located close to the concave side of the cell, but exceeds

Wild type

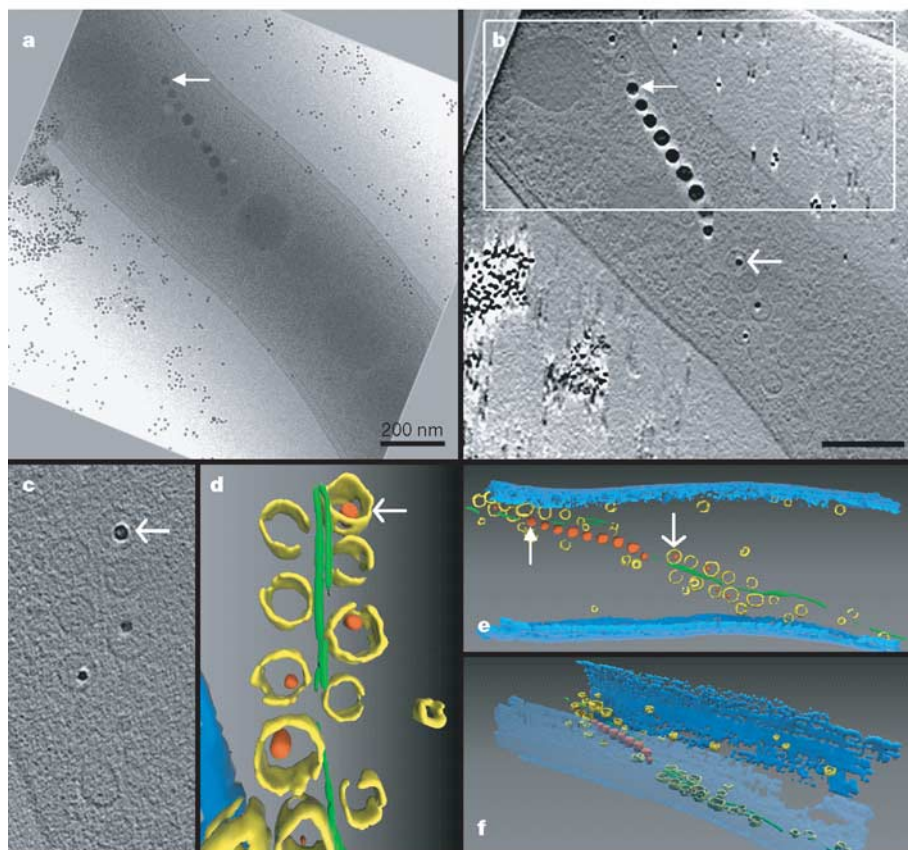
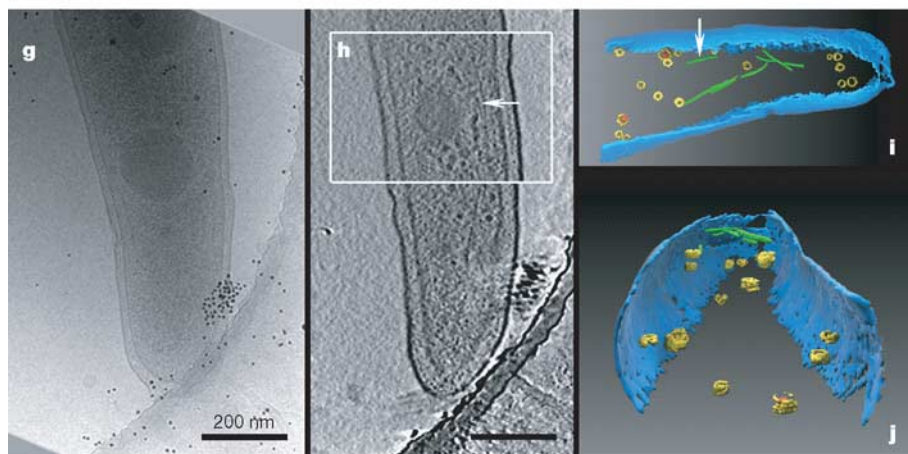


Figure 2 | Cryo-electron tomography of wild-type and ΔmamJ cells. **a–f**, Wild-type cells.

g–j, ΔmamJ cells. Arrows (solid or fine) indicate identical positions in different views. The apparent absence of the magnetosome membrane around mature particles is due to the adverse effect of diffraction contrast of the crystalline magnetite. **a–f**, TEM of a vitrified wild-type cell. **a**, The two-dimensional image represents one projection from an angular tilt-series. **b–f**, Three-dimensional reconstruction of a wild-type cell. **b**, Superimposed x – y slices along the z axis through a typical tomogram. **c**, Magnified x – y slice of vesicles, which are either empty or contain growing crystals (arrow) connected to the filamentous structure. **d**, Surface-rendered representation of a segment of the cell showing vesicles (yellow), magnetite crystals (red) and a filamentous structure (green). **e, f**, Complete three-dimensional visualization of the whole cell in different views (membrane in blue). **g–j**, TEM of a vitrified ΔmamJ cell. **g**, The two-dimensional image represents one projection from an angular tilt-series. **h–j**, Three-dimensional reconstruction of a mutant cell. **h**, Superimposed x – y slices along the z axis through a typical tomogram. **i**, Surface-rendered representation of the whole cell. **j**, In contrast to the wild-type cell, magnetosome vesicles in ΔmamJ cells are scattered throughout the cytoplasm and are not associated with the filaments.

 ΔmamJ 

the length of magnetosome chains, which are usually confined to the midcell. This pattern, which is reminiscent of the localization of a GFP fusion with a different magnetosome protein (MamA)¹², indicates that the localization of MamJ does not depend on the presence of mature magnetite crystals. In contrast, cytosolic localization of the MamJ-EGFP fusion was observed if expressed in the non-magnetic mutant strain MSR-1B, in which a large part of the magnetosome island (including all magnetosome genes) has been deleted^{2,13}. This suggests that MamJ interacts with a linear structure encoded within the magnetosome island, but is not likely to form this structure itself.

To analyse this in greater detail in a close-to-native state, we studied subcellular structures putatively governing the localization of MamJ by means of cryo-electron tomography (cryo-ET), which allows three-dimensional reconstruction at high resolution with a minimum of preparative artefacts. In fully magnetic wild-type cells, mature magnetosome chains were adjacent to the cytoplasmic membrane, apparently following the cell curvature in a helical manner (Supplementary Fig. 2). For better visualization of empty vesicles, we grew cells at moderate iron limitation to reduce the number of magnetite crystals. Tomograms of wild-type and $\Delta mamJ$ cells clearly showed a network of filamentous structures (Fig. 2 and Supplementary Video 1). Individual filaments were approximately

3–4 nm thick, which is in good agreement with similar structures observed in *Spiroplasma melliferum*¹⁴. Filaments extended up to the cell pole (data not shown) and were located close to the cytoplasmic membrane. Notably, most magnetosome vesicles were arranged closely along this filamentous structure in wild-type cells. Empty vesicles and immature magnetosomes were predominantly located at the ends of chains in wild-type cells, and mature magnetosomes were mostly found at midcell. In $\Delta mamJ$ cells, however, empty vesicles and those containing immature crystals were scattered throughout the cytoplasm and were dissociated from the filaments. The organization of the filamentous structures, however, appeared identical to that in wild-type cells. These results indicate that a structure resembling a cytoskeleton is associated with the magnetosome chain, and that the MamJ protein connects magnetosome vesicles to this putative cytoskeletal structure.

In $\Delta mamJ$ cells, empty vesicles were widely spaced and located along the entire length of cells, in contrast to the tightly packed clusters of mature magnetite crystals usually located at midcell. This observation prompted us to examine whether the formation of clusters results from agglomeration of empty magnetosome vesicles, or whether chain collapse is instead caused by magnetic interactions between mature magnetite crystals. We therefore studied the

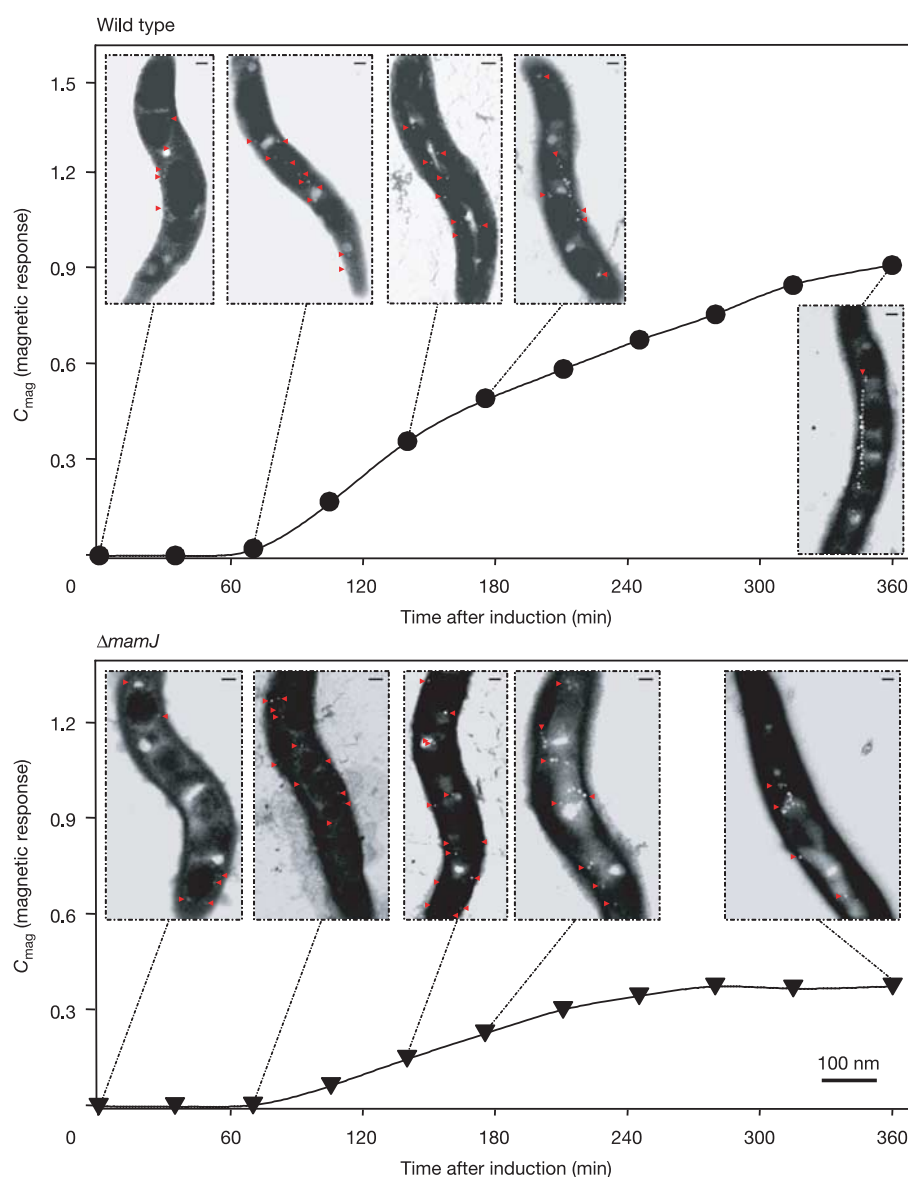


Figure 3 | Time course of magnetite formation in wild-type and $\Delta mamJ$ cells after induction.

Magnetite formation was induced in non-growing iron-deprived cells by the addition of 50 μ M ferric citrate, and was followed by magnetic response measurements (C_{mag}) and TEM. 'Most probable loss images' obtained by energy-filtered TEM show the distribution of growing magnetite crystals (red arrowheads) in characteristic cells after different elapsed times. The micrographs show that growing magnetosome particles seem to undergo dynamic localization during chain assembly, and that the agglomeration of particles in $\Delta mamJ$ cells occurs at a later stage of crystal growth than in wild-type cells. This agglomeration is probably caused by increasing magnetic attraction between particles.

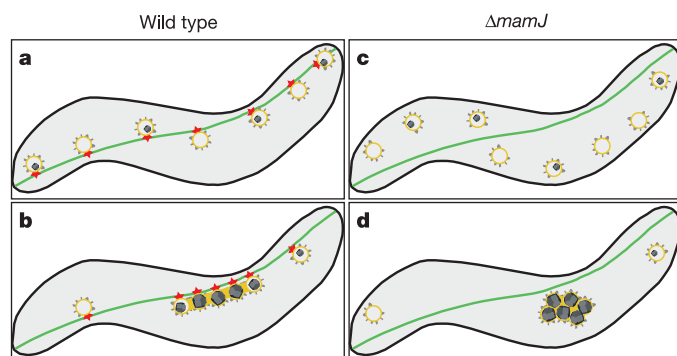


Figure 4 | Model for magnetosome chain assembly. MamJ (red) attaches the empty magnetosome vesicles (yellow) to the cytoskeletal filaments (green), resulting in a loosely spaced 'beads-on-a-string'-like alignment. **a**, Magnetite biomineralization is initiated at multiple discrete sites distributed over the length of the cell. As magnetite crystals (black) continue to grow, their magnetic moments increasingly force the particles to interact. **b**, Positioning of mature magnetosomes at midcell is then directed along filaments. Once particles are in close contact, growing chains are further stabilized by interaction through the magnetosome membrane. **c**, In contrast, vesicles in the $\Delta mamJ$ mutant are detached from the filaments. **d**, Vesicles are free to diffuse within the cells, resulting in agglomeration of particles upon close contact as magnetic interactions between growing magnetite crystals increase.

process of magnetosome formation in a time course experiment (Fig. 3). Approximately 30 min after induction, small (approximately 5–10 nm in diameter) crystallites were formed simultaneously at multiple discrete sites along the entire length of both wild-type and mutant cells. In wild-type cells, growing crystals started to concentrate at midcell after 175 min, first assembling into imperfect, loosely spaced chains that gradually (over 360 min) developed into straight, tightly packed chains of mature particles. At this time, electron-dense particles seemed mostly confined to midcell. This pattern is most likely explained by the dynamic localization of magnetosome particles during chain assembly. In contrast, the localization of immature crystals did not change significantly for about 140 min in $\Delta mamJ$ cells. However, clustering of magnetosomes started at about 175 min, as numerous larger particles were present. This indicates that the agglomeration of particles occurs at a later stage of crystal growth, and is probably caused by increasing magnetic attraction between crystals once they are in close proximity to each other.

Given the essential requirement of MamJ for the assembly of magnetosome chains, we postulate a mechanism that involves magnetic interactions within the chain, physical contact between adjacent particles, and the interaction of MamJ with a cytoskeleton-like structure that directs the assembly and localization of the prokaryotic organelles (Fig. 4). Intriguingly, *mamJ* is co-transcribed with the *mamK* gene (S. Schübbe, C. Würdemann, F. O. Glöckner, and D. S., manuscript in preparation), which has previously been hypothesized to encode a cytoskeletal structure because of its striking similarity to actin-like MreB proteins in other bacteria^{15,16}. Our observation that this cytoskeletal structure seems to be absent in the MSR-1B mutant (which lacks *mamK*) is consistent with this speculated role.

Our results demonstrate that magnetosome chain assembly is genetically controlled to achieve one of the highest structural levels in a prokaryotic cell for optimal function in magnetotaxis. Given that magnetosome chains resembling those in magnetotactic bacteria have been identified in higher organisms^{17,18}, the biological mechanisms of chain formation will be relevant for understanding the evolution and function of magnetoreceptive organelles in both prokaryotic and eukaryotic organisms.

METHODS

Bacterial strains, media and magnetosome isolation. *Escherichia coli* strains were grown as described¹⁹. Liquid cultures of wild-type and mutant strains of *M. gryphiswaldense* MSR-1 (DSM6361) were grown microaerobically in modified FSM medium²⁰. Single colonies were cultivated on ACA medium²¹. Non-magnetic cells were grown at moderate iron limitation by repeated cultivation in modified FSM medium, from which ferric citrate and yeast extract were omitted. In this medium, which contained only trace amounts of iron, cells formed only a few tiny crystallites and did not show a magnetic response. Magnetite induction experiments were performed microaerobically on these non-magnetic cells in modified FSM medium that had a reduced content of carbon sources. Under these conditions, cells did not grow appreciably but remained viable and produced magnetosomes for several hours upon induction of magnetite formation by the addition of ferric citrate to a final concentration of 50 μ M. Magnetosomes were isolated as previously described⁸ from cultures grown microaerobically in an oxygen-controlled fermenter²⁰.

The average magnetic orientation of cell suspensions ('magnetic response') was assayed by an optical method as previously described¹⁰. Briefly, cells were aligned at different angles relative to a light beam by means of an external magnetic field. The ratio of the resulting maximum and minimum scattering intensities (C_{mag}) is well correlated with the average number of magnetic particles and can be used for quantitative assessment of magnetite formation (for practical purposes, non-magnetic cells were assumed to have $C_{mag} = 0$).

Construction and detection of fusion proteins. The *M. gryphiswaldense* *mamJ* gene (accession number AM085146) and *egfp* (pEGFP-N2, Clontech) were amplified using primers ASmamJs_f/ASmamJe_r5 (5'-GGAATTCATGGC AAAAACC GCGGTG-3'; 5'-GCCCTTGCTCACCATT TTTATCTTATCTT CAGC-3') and ASegfp_f1/ASegfp_r3 (5'-GCTGAAGATAAGATAAAATGGT GAGCAAGGGC-3'; 5'-TCTAGATTACTTGTACAGCTCGTCCATGCC-3'), respectively. Using the gene fragments as templates, fusion polymerase chain reaction (PCR) with the ASmamJs_f/ASegfp_r3 primer pair generated a functional 1,998-bp *mamJ-egfp* fusion, which was cloned into the *EcoRI* and *XbaI* sites of the pBBR1MCS2 vector. Expression of MamJ–EGFP in $\Delta mamJ$ cells restored magnetosome chain formation as monitored by electron microscopy. Fluorescence microscopy was performed using an Olympus AX70 microscope equipped with a digital CCD camera. Images were acquired with Metamorph 5.0 (Universal Imaging Corp.) using exposure times of 0.5–5.0 s. Image rescaling and cropping was done using Photoshop 6.0. Cell membranes were stained with FM4-64 (final concentration 1 nM).

Construction of an unmarked in-frame *mamJ* deletion mutant. Gene replacement was performed by homologous recombination of flanking sequence regions of *mamJ* between the *M. gryphiswaldense* chromosome and pAS2. A 1,394-bp upstream region and a 1,433-bp downstream region of *mamJ* were amplified using the primers ASmamJu_f/ASmamJs_r (5'-GGAATCTT ACCCCCAATGGTGC GTTTG-3'; 5'-CCCGGGCATTATCCGCTCCACCCT CAA-3') and ASmamJe_f/ASmamJd_r (5'-CCCGGGTAGAGACCTGTGGATT GATCTGT-3'; 5'-CTCTAGAGCTCATCTTACCTACTCCAAAG-3'), respectively. Fragments were fused by cloning into the *EcoRI* and *XbaI* sites of the pK19mobsacB vector, resulting in pAS2. Plasmids were introduced into *M. gryphiswaldense* by biparental conjugation and screening for recombinants was performed as previously described²². Double-crossover was confirmed by Southern blot analysis. The nonpolar character of the mutation was confirmed by reverse transcription PCR detection of the transcripts of several genes located downstream of *mamJ*. In addition, presence of the product of the *mamK* gene (which is located immediately downstream of *mamJ*) was confirmed using a specific antibody (data not shown). Restoration of the chain-forming phenotype in $\Delta mamJ$ cells was achieved by transcomplementation with wild-type *mamJ* expressed from plasmid pBBR1MCS2.

Electron microscopy. For TEM, cells and isolated magnetosomes were adsorbed on carbon-coated copper grids. Purified magnetosomes were negatively stained with 2% (w/v) uranyl acetate. Bright-field TEM images were obtained with a Zeiss EM10 transmission electron microscope at an accelerating voltage of 60 kV. Energy-filtered TEM (EFTEM) was performed on a Philips CM120 BioTwin equipped with a Gatan energy filter. 'Most probable loss images' were acquired at around 30 eV with a 20-eV energy window and with white and black inverted as described²³. For tomography, wild-type and $\Delta mamJ$ mutant cells were vitrified in liquid ethane and investigated at liquid nitrogen temperature with a Philips CM300 field emission transmission electron microscope equipped with a Gatan energy filter. Multiple single-axis tilt-series at cryo-conditions and in zero-loss imaging mode were recorded and reconstructed by weighted back-projection. All surface-rendered visualizations were de-noised beforehand by anisotropic nonlinear diffusion. Owing to the limited tilt-range ($\pm 60^\circ$), all tomograms

contain artefacts (the 'missing-wedge' effect, that is, information missing perpendicular to the tilt-axis), resulting in surface-rendered representations that show incomplete visualization of the cellular membrane and vesicles.

Received 16 October; accepted 28 October 2005.

Published online 20 November 2005.

1. Bazylnski, D. A. & Frankel, R. B. Magnetosome formation in prokaryotes. *Nature Rev. Microbiol.* **2**, 217–230 (2004).
2. Ullrich, S., Kube, M., Schübbe, S., Reinhardt, R. & Schüler, D. A hypervariable 130 kb genomic region of *Magnetospirillum gryphiswaldense* comprises a magnetosome island, which undergoes frequent rearrangements during stationary growth. *J. Bacteriol.* **187**, 7176–7184 (2005).
3. Bazylnski, D. Structure and function of the bacterial magnetosome. *ASM News* **61**, 337–343 (1995).
4. Bazylnski, D. A., Garratt-Reed, A. J. & Frankel, R. B. Electron-microscopic studies of magnetosomes in magnetotactic bacteria. *Microsc. Res. Tech.* **27**, 389–401 (1994).
5. Dunin-Borkowski, R. E. *et al.* Magnetic microstructure of magnetotactic bacteria by electron holography. *Science* **282**, 1868–1870 (1998).
6. Kirschvink, J. L. Paleomagnetic evidence for fossil biogenic magnetite in western Crete. *Earth Planet. Sci. Lett.* **59**, 388–392 (1982).
7. Vali, H. & Kirschvink, J. L. in *Iron Biominerals* (eds Frankel, R. B. & Blakemore, R. P.) 97–116 (Plenum Press, New York and London, 1991).
8. Grünberg, K. *et al.* Biochemical and proteomic analysis of the magnetosome membrane in *Magnetospirillum gryphiswaldense*. *Appl. Environ. Microbiol.* **70**, 1040–1050 (2004).
9. Evans, J. S. 'Apples' and 'oranges': comparing the structural aspects of biomineral- and ice-interaction proteins. *Curr. Opin. Colloid Interf. Sci.* **8**, 48–54 (2003).
10. Schüler, D., Uhl, R. & Baeuerlein, E. A simple light-scattering method to assay magnetism in *Magnetospirillum gryphiswaldense*. *FEMS Microbiol. Lett.* **132**, 139–145 (1995).
11. Philipse, A. P. & Maas, D. Magnetic colloids from magnetotactic bacteria: Chain formation and colloidal stability. *Langmuir* **18**, 9977–9984 (2002).
12. Komeili, A., Vali, H., Beveridge, T. J. & Newman, D. K. Magnetosome vesicles are present before magnetite formation, and MamA is required for their activation. *Proc. Natl Acad. Sci. USA* **101**, 3839–3844 (2004).
13. Schübbe, S. *et al.* Characterization of a spontaneous nonmagnetic mutant of *Magnetospirillum gryphiswaldense* reveals a large deletion comprising a putative magnetosome island. *J. Bacteriol.* **185**, 5779–5790 (2003).
14. Kürner, J., Frangakis, A. S. & Baumeister, W. Cryo-electron tomography reveals the cytoskeletal structure of *Spiroplasma melliferum*. *Science* **307**, 436–438 (2005).
15. Errington, J. Dynamic proteins and a cytoskeleton in bacteria. *Nature Cell Biol.* **5**, 175–178 (2003).
16. Schüler, D. Molecular analysis of a subcellular compartment: the magnetosome membrane in *Magnetospirillum gryphiswaldense*. *Arch. Microbiol.* **181**, 1–7 (2004).
17. Diebel, C. E., Proksch, R., Green, C. R., Nelson, P. & Walker, M. M. Magnetite defines a vertebrate magnetoreceptor. *Nature* **406**, 299–302 (2000).
18. Mora, C. V., Davidson, M., Wild, J. M. & Walker, M. M. Magnetoreception and its trigeminal mediation in the homing pigeon. *Nature* **432**, 508–511 (2004).
19. Sambrook, J. & Russel, D. W. *Molecular Cloning: A Laboratory Manual* 3rd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001).
20. Heyen, U. & Schüler, D. Growth and magnetosome formation by microaerophilic *Magnetospirillum* strains in an oxygen-controlled fermentor. *Appl. Microbiol. Biotechnol.* **61**, 536–544 (2003).
21. Schultheiss, D. & Schüler, D. Development of a genetic system for *Magnetospirillum gryphiswaldense*. *Arch. Microbiol.* **179**, 89–94 (2003).
22. Schultheiss, D., Kube, M. & Schüler, D. Inactivation of the flagellin gene *flaA* in *Magnetospirillum gryphiswaldense* results in non-magnetotactic mutants lacking flagellar filaments. *Appl. Environ. Microbiol.* **70**, 3624–3631 (2004).
23. Grimm, R., Typke, D., Barmann, M. & Baumeister, W. Determination of the inelastic mean free path in ice by examination of tilted vesicles and automated most probable loss imaging. *Ultramicroscopy* **63**, 169–179 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Graumann for advice on fluorescence microscopy and F. Widdel for helpful comments. This research was supported by the Max Planck Society and the Biofuture program of the Bundesministerium für Bildung und Forschung.

Author Contributions A.S. carried out all genetic and growth experiments and performed fluorescence and TEM microscopy. M.G. carried out cryo-electron tomography and analysis of tomograms. D.F. participated in induction experiments. A.L. participated in three-dimensional visualization. J.M.P. directed cryo-electron tomography, EFTEM experiments and data analysis. D.S. coordinated the study and with A.S. finalized the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.S. (dschuele@mpi-bremen.de).

Development of the signal in sensory rhodopsin and its transfer to the cognate transducer

Rouslan Moukhametzianov^{1,2*}, Johann P. Klare^{3*}, Rouslan Efremov^{1,2}, Christian Baeken¹, Annika Göppner³, Jörg Labahn¹, Martin Engelhard³, Georg Büldt¹ & Valentin I. Gordeliy^{1,2}

The microbial phototaxis receptor sensory rhodopsin II (NpSR_{II}, also named phoborhodopsin) mediates the photophobic response of the haloarchaeon *Natronomonas pharaonis*^{1,2} by modulating the swimming behaviour of the bacterium³. After excitation by blue-green light NpSR_{II} triggers, by means of a tightly bound transducer protein (NpHtr_{II}), a signal transduction chain homologous with the two-component system of eubacterial chemotaxis⁴. Two molecules of NpSR_{II} and two molecules of NpHtr_{II} form a 2:2 complex in membranes as shown by electron paramagnetic resonance⁵ and X-ray structure analysis⁶. Here we present X-ray structures of the photocycle intermediates K and late M (M₂) explaining the evolution of the signal in the receptor after retinal isomerization and the transfer of the signal to the transducer in the complex. The formation of late M has been correlated with the formation of the signalling state^{2,7}. The observed structural rearrangements allow us to propose the following mechanism for the light-induced activation of the signalling complex. On excitation by light, retinal isomerization leads in the K state to a rearrangement of a water cluster that partly disconnects two helices of the receptor. In the transition to late M the changes in the hydrogen bond network proceed further. Thus, in late M state an altered tertiary structure establishes the signalling state of the receptor. The transducer responds to the activation of the receptor by a clockwise rotation of about 15° of helix TM2 and a displacement of this helix by 0.9 Å at the cytoplasmic surface.

The initial event of signal transduction on excitation by light is the isomerization of the retinal chromophore from all-*trans* to 13-*cis*, followed by a series of spectroscopically and kinetically well characterized intermediate states^{8,9}. The transition between M₁ and M₂ ($\tau = 2$ ms)⁸ is thought to encompass the formation of the 'active' signalling state, which persists until the late intermediate O decays to the original ground state^{2,7,10}.

Crystals of NpSR_{II} in complex with shortened transducer molecules (K state NpHtr_{II}₁₋₁₁₄; M state NpHtr_{II}₁₋₁₅₇) were grown in the lipidic cubic phase^{11,12}. The ground-state structure of NpSR_{II} in the complex (Fig. 1a) is almost identical to the structure of NpSR_{II} alone^{13,14}. The crystals of the complex were illuminated to freeze-trap the K and M intermediates. Single-crystal spectrophotometry was used to determine the amount of trapped intermediates (Fig. 1b). Because the M state is split into M₁ and M₂, which cannot be distinguished by their absorption spectra¹⁵, we used flash-photolysis experiments (Fig. 1c), revealing 5% M₁ state and 49% M₂ state (Supplementary Methods). In the K state structural changes with respect to the ground state are observed within a sphere of 9 Å diameter around the central water cluster on the extracellular side of the retinal (Fig. 2a). The isomerization of the retinal bound to Lys 205

gives rise to a displacement of its C ϵ and C δ atoms by 1.1 Å, whereas the backbone atoms show almost no changes from the ground state. This shift reduces the distance between Lys 205-C ϵ and the ground-state position of water molecule 1 (W1-O) from 3.3 to 2.5 Å. Therefore the pentagonal structure of hydrogen bonds including water molecules W1 to W3 and the oxygen atoms of the aspartic residues Asp 75 and Asp 201 found for the ground state can no longer exist in the K state. W1 is sterically driven away from its original location; presumably it shows positional disorder in a cavity formed by the retinal, Asp 75 and Thr 79.

The charge separation between the protonated Schiff-base nitrogen and Asp 75-O δ increases from 4.2 Å to 4.9 Å in the K state. As observed for the K state of the receptor alone¹⁶, the carboxylic group of Asp 75 in the complex is rotated by about 90° from its ground-state conformation, thereby losing the hydrogen bond to Thr 79. Consequently W3 becomes disordered. The hydrogen bond structure (Asp 201-O δ ...W2...W4...Arg 72-Ne) is retained, although W2 and W4 shift towards the retinal by about 1.5 Å in the K state. The distance from Asp 201-O δ to the Schiff-base nitrogen is reduced from 4.9 Å to 2.7 Å as a result of the movement of the Schiff base. The conformation of Arg 72 is not affected by the cascade of changes launched by the photoisomerization of retinal. For all observed photocycle states this residue shows a conformation corresponding to that of Arg 82 in the bacteriorhodopsin M state (Fig. 2c), where it is part of the proton release chain^{17,18}.

During the transition from K state to M₁ state the proton of the Schiff base translocates to Asp 75 (refs 19, 20). Because there is a distance of almost 5 Å in the K state between the Schiff-base nitrogen and Asp 75, the thermal mobility of these groups may be of crucial importance for the proton transfer to succeed. Indeed, during the reaction path from the ground state through K to late M the temperature factor of Asp 75-O δ 2 varies from 18 Å² (B_G) through 31 Å² (B_K) to 22 Å² (B_M) and the distance between Asp 75-O δ 2 and the Schiff-base nitrogen relaxes from 4.9 Å in the K state to 4.3 Å in late M (Fig. 2b), when proton transfer has occurred.

In late M state a further hydrogen bond connecting helices C and G in the K state through water molecules W2' and W4' has vanished (Figs 2b and 3a). Consequently these helices now have more freedom to move independently (Fig. 3; see also stereo view in Supplementary Fig. S1). Difference densities clearly show changes in the position of the main chain that are not present in the K state (for example Arg 72; Fig. 2b). The helical segment between Pro 71 and Pro 81 in helix C is displaced towards the central water cluster by 0.4 Å with respect to the ground state, although the centre of mass of helix C is moving slightly outwards. Water molecule W2' shifts towards Asp 75 and Arg 72 (Fig. 2b).

The main structural changes in late M state are displacements of

¹Research Centre Jülich, Institute of Structural Biology (IBI-2), 52425 Jülich, Germany. ²Centre for Biophysics and Physical Chemistry of Supramolecular Structure, MIPT, 141700 Dolgoprudny, Russia. ³Max Planck Institute for Molecular Physiology, Otto Hahn Strasse 11, 44227 Dortmund, Germany.

*These authors contributed equally to this work.

helices (Fig. 3) that appear after the loss of interhelical connections between helices C and G in the central water cluster (Fig. 2b). Furthermore, the movement of a proton from the Schiff base to Asp 75 alters the charge distribution in this region so that the resulting electrostatic interactions might also influence the tertiary structure. To obtain an overview of all changes in the tertiary structure of the complex we have calculated a double-distance matrix of the difference of the C α atom distances in late M minus the corresponding distances in the ground state (Supplementary Figure S2). This matrix reveals that the receptor is divided into two domains with respect to the relative motions of the helices. The relative motion of helix A with respect to helices B and G is quite small, so these helices are part of the domain ABG. The other domain consists of helices C, D, E and F because double distances between these helices are also quite small (Supplementary Figure S2). Helices in each of the domains are connected by van der Waals contacts along their complete length including several intercalating bulky aromatic side chains. Furthermore, there are several hydrogen bonds between side chains as well as interhelical connections involving water molecules contributing to the rigidity within the domains. The relative movement of these two domains requires structural changes at the interaction surfaces of helices B–C (not shown) and helices F–G (Figs 3b, c and 4a). These interfaces are formed almost exclusively by leucine and isoleucine residues, which allow relative movements (shear motions) of helices by moderate changes of side-chain torsion angles of about 10°. The neighbouring helices B and C of the two domains exert only small movements with respect to each other, whereas helices F and G, the other pair of interdomain helices, have larger relative movements in the direction to membrane normal. Figure 3b, c and 4 show that the main part of this movement

is a shift of helix G towards the extracellular side in late M with respect to helix F. Obviously domain ABG is nearly fixed at the interface between helices B and C and the ABG domain moves with the shift observed at the interface between helices F and G. The relative displacement of helices F and G is stabilized in late M state by the formation of a hydrogen bond between Asp 214 and Arg 162, which might contribute to the longer M₂-state and O-state lifetimes than those of bacteriorhodopsin. It should be noted that the behaviour of helix F is different from that observed for the late M state of bacteriorhodopsin, in which helix F exerted an outward bending motion^{18,21,22}, and also that deduced from electron paramagnetic resonance (EPR)^{5,10} and biochemical²³ studies for NpSRII. This could have two causes: the larger EF loop of bacteriorhodopsin compared with the short loop in NpSRII and/or packing interactions with the adjacent layers in the crystal. The outward movement of helix F in bacteriorhodopsin opens the cytoplasmic domain for the penetration of water molecules, which are necessary for the reprotonation of the Schiff base and Asp 96. In contrast to this, NpSRII, when complexed with its transducer, is not pumping protons^{24,25}, so an opening of the cytoplasmic domain is not required.

Apart from a possible contribution of electrostatic interactions to the signal transfer, which are also effective over larger distances, the signal transfer by short-range interactions takes place in the interface between helices F, G and TM2. All hydrogen bonds between receptor and transducer observed for the ground state are still intact in late M state (Fig. 4a). Our X-ray structures show a constant hydrogen-bond length of 2.6 Å between Tyr 199 (helix G) and Asn 74 (helix TM2) and a parallel shift of both amino acids by about 0.6 Å to the extracellular side in late M. Changes in distance between helix F and TM2 are quite small with the exception of the cytoplasmic side, where a distance

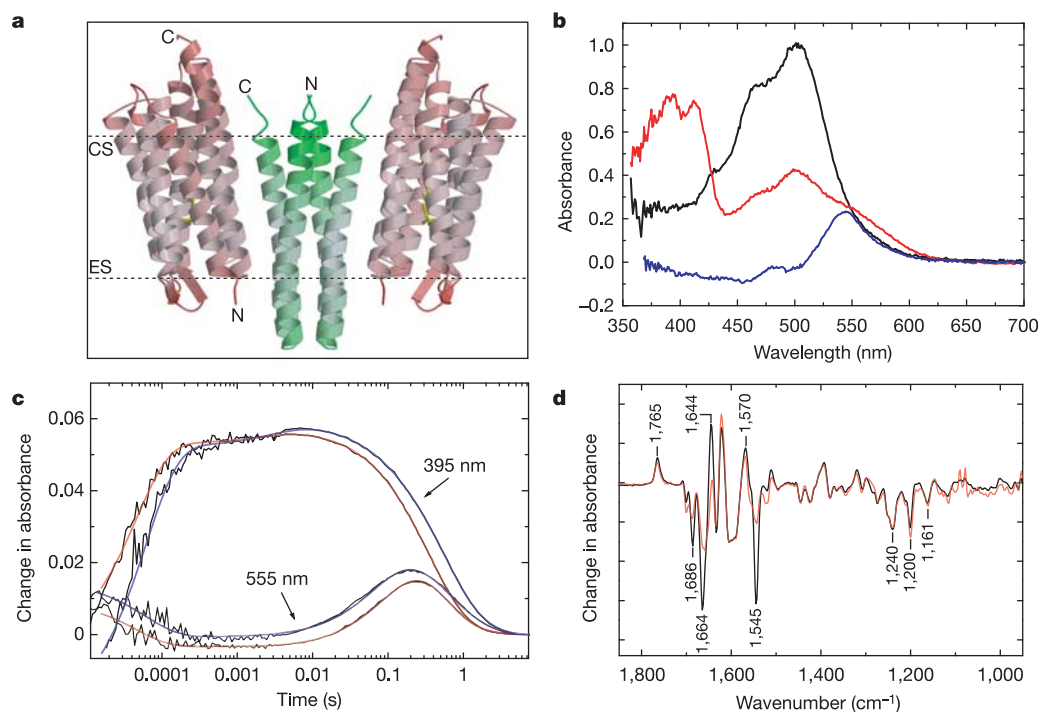


Figure 1 | Crystal structure of the complex and its spectroscopic characterization. **a**, Crystal structure of the complex in the ground state consisting of two molecules of NpSRII (red) with two molecules of NpHtrII (green). The ribbon diagram of the side view of the complex in B-factor colouring: light red and green indicate less mobile; dark red and green indicate mobile. ES, extracellular side; CS, cytoplasmic side. The dotted lines confine the hydrophobic core of the protein. **b**, Low-temperature (100 K) absorbance spectrum of the complex in crystals: ground state (black), trapped M state (red) and a difference spectrum of trapped K state minus

ground state (blue). **c**, Photocycle kinetics of the complex in crystals and of the complex reconstituted in polar lipids from purple membranes of *Halobacterium salinarum* in 3 M sodium phosphate pH 5.6 at 25 °C. Traces of transient absorbance difference of the complex at 390 and 555 nm show the time evolution of the M and O states. Raw data are shown in black, and exponential fits are shown in red and blue for the complex in liposomes and in crystals, respectively. **d**, Difference FTIR spectra of illuminated minus non-illuminated complex in the crystal (red) and reconstituted in liposomes (black) measured at room temperature.

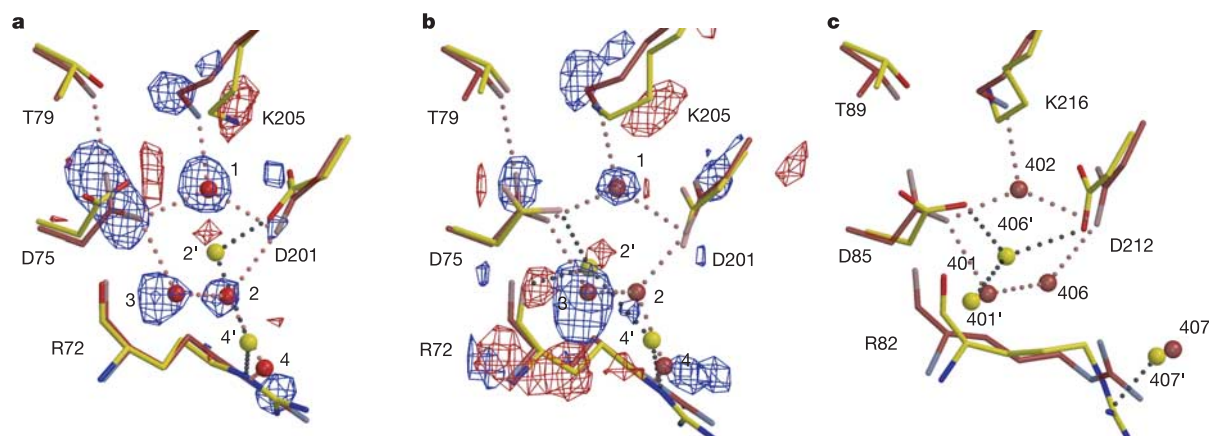


Figure 2 | Structural changes in the extracellular vicinity of the retinal Schiff base. Hydrogen-bond pattern including four water molecules, Lys 205, Asp 75, Asp 201, Arg 72 and Thr 79 side chains, showing ground state (red), intermediate states (yellow), water molecules of ground state (red balls) and intermediate states (yellow balls). Hydrogen bonds are shown as red dotted lines for the ground state and as black dotted lines for the

intermediate states. Difference densities are shown in blue for negative densities and red for positive densities (Methods). **a**, K state of complex with electron density maps contoured at 4.5σ and -4.5σ . **b**, M intermediate of complex with electron density maps contoured at 3σ and -3σ . **c**, Analogue residues of ground-state and late M-state structures of bacteriorhodopsin (Protein Data Bank accession code 1CWQ).

increase of $0.6 \text{ \AA}$ is observed (Fig. 4b, Supplementary Figure S2), for an in-plane movement of TM2. Double distances of helix G and TM2 are larger, so that helix G has an active function in changing the interaction surface faced by TM2 and in this way conducts the signal from the receptor to the transducer. Figures 3c and 4 show that the signal transfer produces a clockwise rotation of TM2 by 15° (observed from the cytoplasmic side) and a tilt of TM2 giving an in-plane displacement of $0.9 \text{ \AA}$ at the cytoplasmic side with the hinge at Gly 66 near Ser 62, which is hydrogen-bonded to Thr 189 (helix G). A rotation of TM2 has already been deduced by EPR experiments of site-directed spin-labelled transducer and receptor molecules⁵. It is striking that the transfer of the signal from the receptor to the transducer is accompanied by an amplification, as reflected by larger amplitudes (rotation and displacements) in the transducer than in the receptor.

Fourier-transform infrared (FTIR) experiments on the complex show a decreased intensity in amide I and II bands in crystals compared with that in liposomes (Fig. 1d), although the localization of bands is conserved, indicating smaller conformational changes in crystals. It may be that the layer packing within the crystal, pre-

dominantly observed between EF and BC loops, restrains the motion of helix F in late M, as mentioned above, or that there are differences in the structure of the region of residues 83–157 that could not be traced.

Thus, light-activated signal transfer from receptor to transducer originates in the isomerization of retinal, which induces local changes in the hydrogen-bonding network in a region near the retinal (K state). During the transition to the signalling state (late M) these disturbances increase so that in late M fewer hydrogen bonds connect helices C and G. These alterations in the hydrogen-bond network change the pK_a values of the Schiff base and the proton acceptor group Asp 75, thereby enabling proton transfer, which produces a redistribution of charges. Both causes give rise to changes in the tertiary structure of the receptor in late M. Signal transfer to the transducer takes place within the interface of receptor helices F and G and transducer helix TM2. The main effect on TM2 is a clockwise rotation of about 15° and a tilt of the helix, with the hinge at Ser 62, that amounts to $0.9 \text{ \AA}$ at the cytoplasmic side (Leu 77). The cytoplasmic end of TM2 rolls over the surface from helix F to G, whereas the extracellular side of TM2 remains fixed. There is a

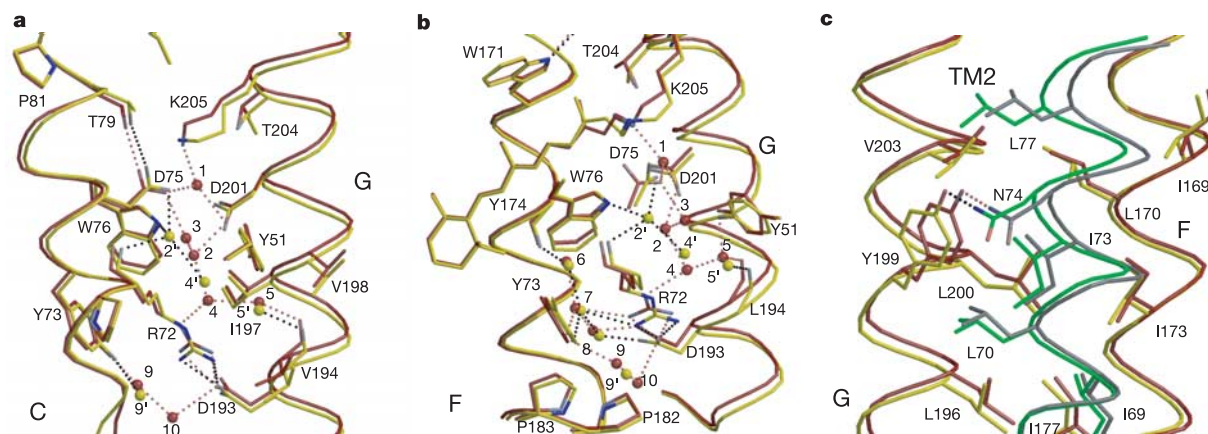


Figure 3 | Interhelical interactions. Changes in helical structure and interactions between ground state (red) and late M state (yellow) including water molecules as red and yellow balls; the hydrogen-bonded network is

shown in red for the ground state and in black for late M state. **a**, Helices C and G; **b**, helices F and G; **c**, helices F, G and TM2 (transducer helix in grey for ground state and in green for late M state).

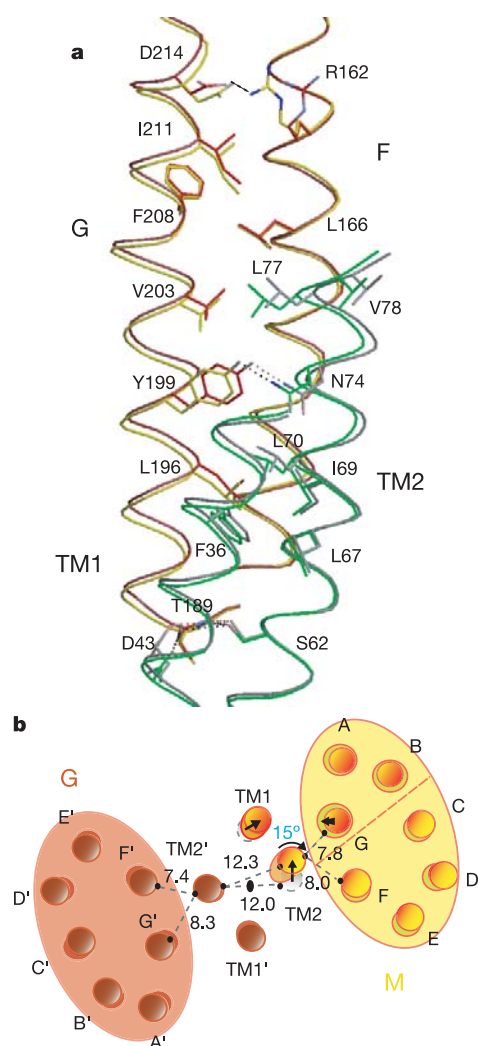


Figure 4 | Overall views of helices F, G, TM1 and TM2 in ground and late M states. **a**, Structural changes along these helices; colours are as in Fig. 3c. **b**, Schematic picture of helical displacements viewed from the cytoplasmic surface. The ground-state complex is shown in red and the late M-state complex in yellow; black numbers give distances between C α atoms of the following amino acids in the ground and late M states: F(Leu 170) to TM2(Leu 77), ground state 7.4 Å, M state 8.0 Å; G(Val 203) to TM2(Leu 77), ground state 7.8 Å, M state 8.3 Å; TM2(Val 78) to TM2' (Val 78), ground state 12.0 Å, M state 12.3 Å. The arrows on TM1 and TM2 depict displacements parallel to the membrane plane in late M; the thick arrow on helix G depicts its 0.5 Å movement to the extracellular side. A clockwise rotation of 15° is observed for TM2 and an in-plane displacement of 0.9 Å (black arrow) near the cytoplasmic membrane surface in late M state with respect to the ground state. A red line in the receptor coloured yellow divides the molecule into the two functionally important domains ABG and CDEF. The inner red shaded circles of helices A, B and G depict their movements to the extracellular side.

combination of rotation and tilt; the latter appears as an in-plane displacement of the cytoplasmic end of TM2.

METHODS

Protein preparation and crystallization. Protein preparation and crystallization were performed as described previously^{5,6}.

Trapping of intermediates. To obtain the highest yield in the occupancy of the K and M states we always used the optimum wavelength and intensity of an argon ion laser. The K state was accumulated by illumination of the crystal at 100 K for 20 s with the argon laser (472 nm, about 10 mW mm⁻²). For trapping late M state the crystal was left to warm up to about 20 °C (by blocking the nitrogen cryostream), illuminated for 3 s with argon laser light at 488 nm (about

30 mW mm⁻²), then cooled (by unblocking the cryostream) while the blue light was still on. One second after recooling started, the illumination was turned off. X-ray data were collected in darkness.

Spectroscopic characterization of the crystals. Visual and FTIR spectra of single crystals (Fig. 1b–d) were measured with an infrared microscope described elsewhere¹². Visual absorption spectra (Fig. 1b) were recorded from single crystals in the nitrogen cryostream under conditions identical to those used for X-ray measurements.

Treatment of diffraction data. X-ray diffraction data (wavelength 0.934 Å) were collected at beamlines ID14-1 and ID13 of the European Synchrotron Radiation Facility (ESRF), Grenoble, France. Low-resolution and high-resolution data sets (1° oscillation) were obtained from several crystals (for ground state seven crystals were used; for K state three were used; for late M state five were used). Diffraction pattern were processed with MOSFLM/SCALA (CCP4). To compare the quality of the different data sets a range of R_{merge} was determined by calculating R_{merge} between pairs of data sets (ground state, 0.07–0.09; K state, 0.10–0.12; M state, 0.15–0.17; ground with M state, 0.22–0.25). The first three values show that the difference in diffraction for each state is in a narrow range whereas the fourth value shows the overall change between the ground and M states. Crystallographic parameters and refinement statistics are summarized in Supplementary Table 1 for two of the best crystals, one crystal for the K state and one larger crystal for the ground and M states using different spots for the same crystal. The mean $[I/\sigma I]$ value for the ground state is smaller than for the M state, as a result of the weak reflections at the high-resolution shells. When calculating mean $[I/\sigma I]$ for the ground state at the M-state resolution of 2.2 (2.26–2.2) Å, values of 8.0 (5.6) were obtained. In our refinement, doubtful reflections were rejected with the strong criterion $F/\sigma F < 2$, resulting in a smaller number of reflections used in the refinement (Supplementary Table 1).

Refinement. Starting from a polyaniline model of NpSR/II/NpHtrII(1–114) (Protein Data Bank accession code 1H2S; ref. 6) the molecular replacement (MOLREP) solution was completed by the automated refinement procedure (ARP)²⁶ using ground state data. Residues 23–82 of the transducer were built. Simulated annealing and individual B -factor refinement were used (CNS)²⁷. Several data sets from non-illuminated crystals gave almost identical models of the ground state with a mean coordinate variation of 0.1 Å, which fits with σ_A values in the K and late M states of 0.16 and 0.12 Å, respectively.

Difference density maps $\Delta\rho = (|F_{\text{obs}}| - |F_{\text{Gobs}}|)\exp(i\phi_G)$ revealed major structural changes that occurred on illumination of the crystals (Fig. 2a, b), with $|F_{\text{obs}}|$ being the observed amplitudes from illuminated crystals, $|F_{\text{Gobs}}|$ the observed ground-state amplitudes and ϕ_G the phases calculated from the ground-state model. Refinements of intermediate states were performed by starting with a superposition of a fixed and a refinable ground-state structure with occupancies varying in the range 0.2–0.8 by simulating annealing (using 3,000 and 5,000 K and a slow cooling protocol in CNS). Extrapolated densities were calculated for refined models $\rho_{\text{ex}} = (1 - \alpha)|F_G|[\exp(i\phi_G)] + \alpha|F_X|[\exp(i\phi_X)]$ with different occupancies α for the intermediate state and correlated with observed omit densities $\rho_{\text{omit}} = |F_{\text{obs}}|[\exp(i\phi_{\text{omit}})]$, with $|F_G|$ being the amplitudes of the ground-state model, $|F_X|$ and ϕ_X being the amplitudes and phases of the intermediate-state models for the K state or late M state, and ϕ_{omit} being the phases calculated from the ground-state model where atoms are omitted. Intermediate state occupancies were estimated from the best correlation as $40 \pm 10\%$ for the K state and $50 \pm 10\%$ for late M state. This variation had no significant effect on the conformation for the intermediate state after refinement. After occupancy had been fixed, models were further refined and manually rebuilt in program O until convergence of R factors was reached. The significance of structural changes was also proven by Supplementary Fig. S3.

Spectroscopic analysis of crystals trapped in the K state contained about 50% K state and 50% ground state so that a refinement by two models for ground and K states is justified. Crystals trapped in the M₂ state contained small amounts of the O, M₁ and K states. The quantity in the M₁ state was estimated to be about 5% and that in the K state even lower. A refinement of crystals trapped in late M by a ground-state and M₂-state model is therefore justified because traces of intermediates less than 10% cannot be detected at our resolution.

Received 25 October; accepted 15 December 2005.

Published online 1 February 2006.

- Spudich, J. L. Variations on a molecular switch—transport and sensory signalling by archaeal rhodopsins. *Mol. Microbiol.* **28**, 1051–1058 (1998).
- Klare, J. P. *et al.* The archaeal sensory rhodopsin II/transducer complex: a model for transmembrane signal transfer. *FEBS Lett.* **564**, 219–224 (2004).

3. Scharf, B. E. & Wolff, E. K. Phototactic behaviour of the archaeobacterial *Natronobacterium pharaonis*. *FEBS Lett.* **340**, 114–116 (1994).
4. Rudolph, J. & Oesterhelt, D. Deletion analysis of the *che* operon in the archaeon *Halobacterium salinarum*. *J. Mol. Biol.* **258**, 548–554 (1996).
5. Wegener, A. A., Klare, J. P., Engelhard, M. & Steinhoff, H.-J. Structural insights into the early steps of receptor–transducer signal transfer in archaeal phototaxis. *EMBO J.* **20**, 5312–5319 (2001).
6. Gordeliy, V. I. *et al.* Molecular basis of transmembrane signalling by sensory rhodopsin II–transducer complex. *Nature* **419**, 484–487 (2002).
7. Yan, B., Takahashi, T., Johnson, R. & Spudich, J. L. Identification of signaling states of a sensory receptor by modulation of lifetimes of stimulus-induced conformations: The case of sensory rhodopsin II. *Biochemistry* **30**, 10686–10692 (1991).
8. Chizhov, I. *et al.* The photophobic receptor from *Natronobacterium pharaonis*: Temperature and pH dependencies of the photocycle of sensory rhodopsin II. *Biophys. J.* **75**, 999–1009 (1998).
9. Hirayama, J. *et al.* Photocycle of phoborhodopsin from haloalkaliphilic bacterium (*Natronobacterium pharaonis*) studied by low-temperature spectrophotometry. *Biochemistry* **31**, 2093–2098 (1992).
10. Wegener, A. A., Chizhov, I., Engelhard, M. & Steinhoff, H.-J. Time-resolved detection of transient movement of helix F in spin-labelled *Pharaonis* sensory rhodopsin II. *J. Mol. Biol.* **301**, 881–891 (2000).
11. Landau, E. M. & Rosenbusch, J. P. Lipidic cubic phases: A novel concept for the crystallization of membrane proteins. *Proc. Natl Acad. Sci. USA* **93**, 14532–14535 (1996).
12. Gordeliy, V. I., Schlesinger, R., Efremov, R., Büldt, G. & Heberle, J. in *Membrane Protein Protocols: Expression, Purification, and Crystallization* (ed. Selinsky, B.) 305–316 (Humana Press, Totowa, New Jersey, 2003).
13. Royant, A. *et al.* X-ray structure of sensory rhodopsin II at 2.1-Å resolution. *Proc. Natl Acad. Sci. USA* **98**, 10131–10136 (2001).
14. Luecke, H. *et al.* Crystal structure of sensory rhodopsin II at 2.4 angstroms: Insights into colour tuning and transducer interaction. *Science* **293**, 1499–1503 (2001).
15. Balashov, S. P., Sumi, M. & Kamo, N. The M intermediate of *Pharaonis* phoborhodopsin is photoactive. *Biophys. J.* **78**, 3150–3159 (2000).
16. Edman, K. *et al.* Early structural rearrangements in the photocycle of an integral membrane sensory receptor. *Structure* **10**, 473–482 (2002).
17. Luecke, H., Richter, H.-T. & Lanyi, J. K. Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science* **280**, 1934–1937 (1998).
18. Sass, H. J. *et al.* Structural alterations for proton translocation in the M state of wild-type bacteriorhodopsin. *Nature* **406**, 649–653 (2000).
19. Engelhard, M., Scharf, B. E. & Siebert, F. Protonation changes during the photocycle of sensory rhodopsin II from *Natronobacterium pharaonis*. *FEBS Lett.* **395**, 195–198 (1996).
20. Furutani, Y., Iwamoto, M., Shimono, K., Kamo, N. & Kandori, H. FTIR Spectroscopy of the M photointermediate in *pharaonis* phoborhodopsin. *Biophys. J.* **83**, 3482–3489 (2002).
21. Koch, M. H. *et al.* Time-resolved X-ray diffraction study of structural changes associated with the photocycle of bacteriorhodopsin. *EMBO J.* **10**, 521–526 (1991).
22. Subramaniam, S., Gerstein, M., Oesterhelt, D. & Henderson, R. Electron diffraction analysis of structural changes in the photocycle of bacteriorhodopsin. *EMBO J.* **12**, 1–8 (1993).
23. Yoshida, H., Sudo, Y., Shimono, K., Iwamoto, M. & Kamo, N. Transient movement of helix F revealed by photo-induced inactivation by reaction of a bulky SH-reagent to cysteine-introduced *pharaonis* phoborhodopsin (sensory rhodopsin II). *Photochem. Photobiol. Sci.* **3**, 537–542 (2004).
24. Bergo, V. B., Spudich, E. N., Rothschild, K. J. & Spudich, J. L. Photoactivation perturbs the membrane-embedded contacts between sensory rhodopsin II and its transducer. *J. Biol. Chem.* **280**, 28365–28369 (2005).
25. Schmies, G., Engelhard, M., Wood, P. G., Nagel, G. & Bamberg, E. Electrophysiological characterization of specific interactions between bacterial sensory rhodopsins and their transducers. *Proc. Natl Acad. Sci. USA* **98**, 1555–1559 (2001).
26. Perrakis, A. *et al.* Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
27. Brunger, A. T. *et al.* Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K.-E. Jaeger and H. Gieren for their help in expressing NpSRII; G. Kachalova, J. Granzin and J. Heberle for scientific discussions; B. Gehrmann for administrative assistance; the staff of beamlines ID14-1 and ID13, in particular E. Mitchell; and the staff of beamline X13 at DESY, in particular W. R. Rypniewski. This study was supported by the Deutsche Forschungsgemeinschaft, the Max-Planck-Gesellschaft, the Alexander von Humboldt Foundation and the Helmholtz Association of National Research Centres.

Author Information The coordinates for these structures have been deposited in the Protein Data Bank under accession codes 2F93 and 2F95. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.B. (g.bueldt@fz-juelich.de) or M.E. (martin.engelhard@mpi-dortmund.mpg.de).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

China syndrome

The US National Science Foundation's *Science and Engineering Indicators 2006* could perhaps be renamed 'Here Comes China'. The biennial report shows an increasingly international science and technology workforce, with China showing large gains in internal investment in R&D, investment by multinational corporations, and numbers of Chinese nationals earning science and engineering doctorates in the United States.

China has increased its R&D investment 24% per year over the past five years, compared with 4–5% for the United States. This growth, from US\$12.4 billion in 1991 to \$84.6 billion in 2003, puts the country behind only Japan and the United States. Meanwhile, investment by US-based multinationals into Asian markets outside Japan has more than doubled, from \$1.5 billion in 1994 to \$3.5 million in 2002, with more than \$1 billion going into China alone. Finally, Chinese students earn more US science and engineering PhDs than those of any other foreign nation.

These statistics are impressive, but they tell only one side of the story. What do they mean in terms of jobs and who will get them? The United States, Europe and Japan still

produce many PhDs and create a host of jobs. But China is coming on strong. One wild card is whether Chinese PhDs will stay in the United States or return home. While China's PhD production in the United States has increased, PhDs by US white males has dropped from its peak of about 8,900 in 1994 to just over 7,000 in 2003.

It would be premature to say this marks the end of US dominance in science and engineering employment, but it does show that the United States is producing less of its own scientists and may have more difficulty recruiting from abroad as other nations, particularly China, ramp up funding and infrastructure. As the report says, these trends point to a "potentially diminished US success in the increasing international competition for foreign scientists and engineers".



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building,
 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)
Natureevents: Silke Opstrup (4994)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
**Germany/Austria/
 The Netherlands:**
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

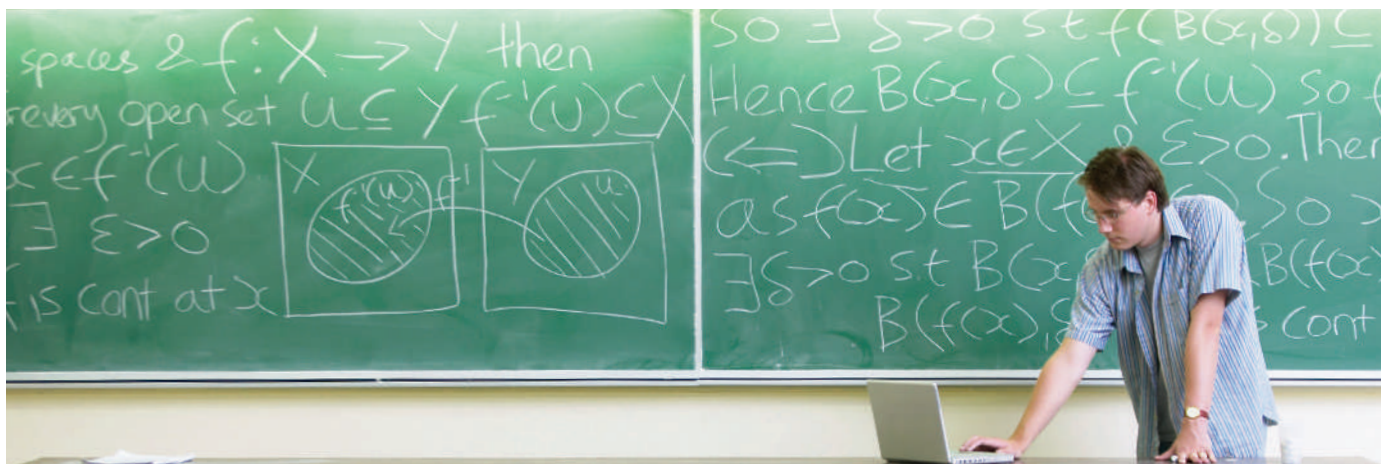
**European Satellite Office
 Germany/Austria/
 The Netherlands:**
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com



P. M. FISHER/CORBIS

After-school programmes

PhD students, postdocs and even senior scientists are taking continuing-education courses to improve their scientific 'hard skills' or branch out beyond the lab.

Robert Rentzsch had a look around to see what's on offer.

Rod MacKinnon was already an established lab head when he decided he needed to learn more about crystallography. In 1995 he went back to school at the Cold Spring Harbor research laboratory (CSHL) on Long Island, New York. Together with 22 fellow scientists, ranging from PhD students and postdocs to principal investigators like himself, he enrolled in a two-week course in macromolecular crystallography. The training paid off: three years later, MacKinnon became the first scientist to use crystallography to resolve the structure of an ion channel. In 2003 he won the Nobel Prize in Chemistry for this work.

MacKinnon's experience may not be typical for the scientists participating in the postgraduate training courses that are offered by a handful of institutes around the world. But it shows that scientists can benefit from acquiring new skills at any stage of their career. Courses range from two to six weeks in length and focus on the latest tools and technologies in a specific field. Classes are kept relatively small, usually about 20 people, providing students with the opportunity to get hands-on experience in well-equipped labs between lectures and even conduct their own mini-projects.

"You could really work as much as you liked," says Olivier Keech, a 26-year-old protein researcher at the Umeå Plant Science Centre in Sweden, who last year participated in the two-week course on proteomics held at the CSHL.

"If you wanted to do your own experiments at two in the morning there was no problem finding some people happy to help you and machines ready to use," adds Keech. He says that he has benefited from the course, and now uses the methods he learnt there in his day-to-day work.

The courses feature "a lot of high tech, like mass specs and robots," says David Stewart, the CSHL's director of meetings and courses. And it left a big mark

on Keech. "Everything was perfectly organized, from food to the invited speakers," he says.

To rustle up the \$3,500 fee and pay for their travel, many students have to do some fundraising. Keech obtained one third of the money from a Swedish grant, another third from a CSHL stipend and the rest from his home institute.

Another hotspot for such courses is the Marine Biological Laboratory (MBL) in Woods Hole, Massachusetts. The courses and lectures here include all the latest methods used in laboratory biology, with a focus on microbiology and neuroscience.

Putting theory into practice

Anke Borgmann, who is doing a PhD in neurobiology at the University of Cologne, Germany, had clear expectations before she booked her flight to Massachusetts. "I studied biology and mathematics, and I was keen to learn how to use my theoretical maths knowledge for my thesis in biology in a gainful way," she says.

The MBL course Borgmann took in computational neuroscience more than met her wishes. "It really changed a lot for me," she says. "Before, I had only read in papers the names of people such as Bard Ermentrout or Larry Abbott. Imagine what it is like to meet them face-to-face."

In the final week, when students were applying what they had just learnt to their own projects, Borgmann remembers working sessions lasting late into the night. But that was fun, she says, not a burden. And so was the faculty dinner, where she had the chance to meet all the professors, assistants and fellow students in a relaxed atmosphere. "I don't think I would have found something comparable in Germany," she says.

Most courses such as those offered by the CSHL, the MBL or the Jackson Laboratory in Maine are heavily oversubscribed. Only about a quarter of applicants are accepted each year. Course leaders say that personal



Olivier Keech: could get help in the lab at 2 a.m.

motivation and the degree to which the applicant would benefit from the training count more than experience and a glossy CV.

In Europe, some of the most notable training courses are those offered by the European Molecular Biology Organization (EMBO). Yannick Schwab, a junior scientist at the Institute of Genetics and Molecular and Cellular Biology in Strasbourg, France, last year attended a two-week EMBO course on electron microscopy in cell biology in the small town of Ceské Budejovice, Czech Republic. Schwab had just been offered a short-term position in Strasbourg on a project aimed at implementing electron microscopy in immune-cell labelling studies.

"It was exactly what I needed at this point," he says. "I now spend 95% of my time using the techniques I learnt in the course." Having experienced teachers available all day to answer questions was a big advantage, says Schwab. "I learnt all the tricks which otherwise would take you 20 years of learning by mistakes."

Sharing new-found knowledge

EMBO is keen that participants should pass on their newly acquired skills to others at their home institutes, as Schwab has been busy doing in Strasbourg. He gives briefings at the local electron-microscopy facility.

EMBO's courses are not limited to Europe. Other recent microscopy courses were held in Santiago, Chile, in 2003, and in Cape Town, South Africa, last year — among many others on its world programme.

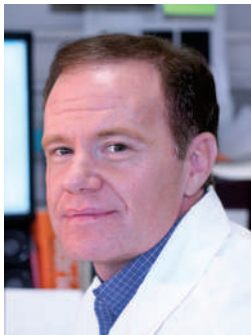
A mixture of European and international participants, mostly postdocs and PhD students, attend the EMBO courses, which are free, except for travel costs. A letter describing personal motivation is important for a successful application, says Gareth Griffith, group leader of the EMBO cell-biology programme.

Those who want to keep up work while digging into something new might head to England, where the University of Oxford's part-time bioinformatics programme leads to a certificate in one year, a diploma in two or a master's degree in three. Participants, mostly professionals, are usually sent by their employers, but individual scientists are also part of the programme. They come from a variety of backgrounds: biology, mathematics or computer science. And most of the coursework can be done from home — participants need only be in Oxford for three or four weeks per year. The course, however, comes at a price. The three-year MSc programme costs European citizens £17,500 (US\$30,500), and those from elsewhere £24,300.

Tom Chittenden, a postdoc at Dartmouth Medical School in Hanover, New Hampshire, has just graduated from the three-year programme. He admits it was difficult balancing his job with the training. "Luckily, I had a very understanding boss," he says. "When writing my thesis last summer I almost faced a nervous breakdown. But now that I'm done it opens a lot of doors to me."

Public-health officials and researchers who want to learn new analytical methods could try a course offered by Imperial College London. It teaches state-of-the-art computer modelling using data from real events, such as the 2003 outbreak of SARS in Hong Kong or the current spread of avian flu.

Although modelling involves a lot of maths and



Tom Chittenden: hard to balance job with training.



Katherine Moortgat: going from idea to IPO.

statistics, there is no need to be afraid, say Peter White, the course organizer, who notes that most people on the course don't do maths at all in their jobs. "We use easy-to-learn modelling tools that do the number stuff for you," he says. The next course will be held this September.

Scientists seeking a bigger career switch into finance, consulting or entrepreneurship have a growing number of business schools interested in tailoring or creating MBA programmes for them. Interesting examples have evolved in the California biotech hubs of San Diego and San Francisco, where business skills are in demand.

"Biotech started on our doorstep, so it's no wonder we get excellent speakers from industry," says Katherine Moortgat, director of the Center for Bioentrepreneurship at the University of California, San Francisco (UCSF).

In one of the centre's classes aimed at budding entrepreneurs, called "Idea to IPO... and Beyond", industry leaders explain the nuts and bolts of building a successful business. Team projects and internships provide hands-on experience. Students write business plans and, at the end of the three-month course, pitch their ideas and results to a panel of venture capitalists.

In San Diego, a group of scientists, still working in their day jobs, meets every other week on Fridays and Saturdays as part of the 24-month 'flexMBA' programme offered by the University of California, San Diego (UCSD). Most applicants for the \$36,000 programme already hold a PhD in science or technology and have a few years of professional experience.

"This is no simulation. It's real work and real stuff," says JoAnne Starr, assistant dean of MBA programmes at UCSD's Rady School of Management. Two of the workshop projects have already received internal funding by companies and a third is in the process of raising venture capital.

Start-ups and job offers

Niall O'Donnell, who did a PhD in protein biochemistry at Oxford and a postdoc at UCSD, will graduate from this programme later this year. "I was really wet behind the ears before. I knew nothing," he says. Recently he got a job offer from a venture-capital firm in Palo Alto, California, partly because he was involved in creating two start-ups during his studies. He has no worries about going home one day. "In Britain you get no job with a PhD nowadays, but with an MBA that looks different," he says.

Similar programmes are being established in other parts of the world as well. Danube University in Krems — Austria's first postgrad-only university — has recently started a two-year part-time MBA. It was inspired by successful schemes such as those offered at the Kellogg School of Management in Chicago and the Keck Graduate Institute in Claremont, California. A new Master of Bioscience Enterprise programme will also start later this year at the University of Auckland in New Zealand.

Whether in the States, in Europe or the southern hemisphere, training can be an inspiring break from scientific routine. And if you're lucky it can be great fun as well. Keech has happy memories of his course at Cold Spring Harbor. "After work we certainly went to the pub for a few drinks," he says.

Robert Rentzsch recently completed an internship in Nature's Munich office.

Applied mathematical theology

You have a message.

Gregory Benford

The discovery that the cosmic microwave background has a pattern buried in it unsettled the entire world.

The temperature of this 2.7 K emission, left over from the Big Bang, varies across the sky. Temperature ripples can be broken into angular-coordinate Fourier components, and this is where radio astronomers found something curious — a message or, at least, a pattern. Spread across the microwave sky there was room in the detectable fluctuations for about 100,000 bits — roughly 10,000 words.

Although different technical civilizations in our Universe would see different temperature fluctuations, they could agree on the Fourier coefficients. This independence of place, and the role of the cosmic background as cosmic neon sign for anyone with a microwave receiver, meant that any intelligence in the Universe could see this pattern.

But what did it mean? Certainly it would not be in English or any other human language. The only candidate tongue was mathematics.

Writing them as binary numbers, astronomers tried to fit mathematical sequences, such as the prime numbers, in any base. This and other mathematical favourites — π , e , the golden ratio, the Riemann zeta function — proved futile. More obscure numbers and patterns, from set theory and the like, also shed no light.

In despair, some thought the pattern might be random. But the Shannon entropy test showed clear non-random elements, and this nihilist idea faded away. One insight from Benford's law, which states that the logarithms of artificial numbers are uniformly distributed, did apply to the tiny fluctuations. This proved that the primordial microwaves were not random, and so had been artificially encoded, perhaps by some even earlier process. So there was a message, of sorts.

Cosmologists eagerly searched for clues and hit a dead end. The sequence was found to fit no model. This suggested immediately to even non-religious astronomers that the pattern might have been put there by a being who made our Universe: God, in short.

What would such a mathematical message mean, anyway? Only that some rational, counting designer had made our Universe. Beyond that, nothing would be revealed about the being's nature; although of course this proved the old claim, that God was a mathematician.

Rankled, the physicists quickly compared the observed sequence with the fine-structure constant, one of their favourites. The sequence did not fit.

This sent everyone back to fundamen-

easier for physicists? Because, obviously, God was one, too.

Sadly, no. Nothing seemed to work.

Perhaps the very idea underpinning science — that humans could understand the Universe — had hit a wall. This helped both science and religion.

Excitement increased. If the being was not saying something obvious, then maybe humans had not understood the Universe well enough to make out the message. Governments poured money into mathematics

and physics. The astronomers protested. If the night sky was a tale told by God, they could read it. The cosmic-neutrino and gravity-wave backgrounds had not yet been detected, but they could also carry the Word. So it came to be that the cosmologists, too, received the blessing of a large research bounty.

These huge increases in funding drove a renaissance of modern science. Data processors, statistical theorists, observers of obscure spectra — all received a share. Vast telescopes tuned to the vibrations and emissions of

the Universe glided in high orbits, their ears cupped to the distant and primordial.

This largesse produced an economic boon, as many spin-off technologies benefited commerce. Religious fervour damped, as each faith felt humbled by this proof that the Universe had meaning, yet mankind was not yet advanced enough to fathom it.

At the same time, attention focused on the injunction to mankind in the Old Testament — echoed in other religious founding texts — charging humanity with being the stewards of Earth. The environmental movement merged with the great religions.

Within a century, active adjustment of Earth's reflected sunlight, and capturing of carbon in the oceans and lands, had averted the greenhouse disaster. Church attendance was enormous. Efforts to enhance our knowledge and skills had averted many gathering social conflicts.

Work on the message continues in the new university departments of applied mathematical theology. Yet to this day, it remains untranslated. Perhaps that is just as well.

Gregory Benford is a professor of physics at the University of California, Irvine. His best-known novel is *Timescape*.



JACEY

tals. Current theory said that tiny temperature fluctuations in the microwaves came from little bumps in the potential function that governed the inflation of the very early Universe. Tinkering with those quantum fluctuations, a being could write something simple but profound: God as a quantum mechanic. If, for example, the designer could encode little squiggles on the potential, then the fine-tuned primordial density fluctuations would not be exactly scale-free, and that's where the sky-wide microwave patterns came from.

So of course the physicists followed their current fashion. When comparison with other favourite numbers — the dimensionless ratios of masses and energies and the like — all failed, they tried more advanced theories. They tried prescriptions for various symmetry groups that came from the Lie algebras, as three of the four fundamental interactions we know reflect such gauge theories. No help.

The physicists, who had long been the mandarins of science, then supposed that clues to the correct string theory, a menu currently offering about 10^{100} choices, would be the most profound of messages. After all, wouldn't God want to make life